Text as Data

Justin Grimmer

Associate Professor
Department of Political Science
Stanford University

November 10th, 2014

Supervised Learning Methods

1) Task
   - Classify documents to pre existing categories
   - Measure the proportion of documents in each category
2) Objective function
   1) Penalized Regressions
      - Ridge regression
      - LASSO regression
   2) Classification Surface $\rightsquigarrow$ Support Vector Machines
   3) Measure Proportions $\rightsquigarrow$ Naive Bayes(ish) objective
3) Optimization
   - Depends on method
4) Validation
   - Obtain predicted fit for new data $f(\boldsymbol{X}_i, \widehat{\boldsymbol{\theta}})$
   - Examine prediction performance $\rightsquigarrow$ compare classification to gold standard

Regression models

Suppose we have N documents, with each document i having label $y_i \in \{-1, 1\} \rightsquigarrow \{\text{liberal, conservative}\}$

Regression models

Suppose we have N documents, with each document i having label $y_i \in \{-1, 1\} \rightsquigarrow$ {liberal, conservative}

We represent each document i is $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$.

Regression models

Suppose we have N documents, with each document i having label $y_i \in \{-1, 1\} \rightsquigarrow \{\text{liberal, conservative}\}$
We represent each document i is $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$.

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^{N} \left(y_i - \boldsymbol{\beta}^{'} \boldsymbol{x}_i\right)^2$$

Regression models

Suppose we have N documents, with each document i having label $y_i \in \{-1, 1\} \rightsquigarrow \{\text{liberal, conservative}\}$
We represent each document i is $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$.

$$
\begin{aligned}
f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) &= \sum_{i=1}^{N} \left(y_i - \boldsymbol{\beta}^{'} \boldsymbol{x}_i\right)^2 \\
\widehat{\boldsymbol{\beta}} &= \text{arg min}_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} \left(y_i - \boldsymbol{\beta}^{'} \boldsymbol{x}_i\right)^2 \right\}
\end{aligned}
$$

Regression models

Suppose we have N documents, with each document i having label $y_i \in \{-1, 1\} \rightsquigarrow$ {liberal, conservative}
We represent each document i is $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$.

$$
\begin{aligned}
f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) &= \sum_{i=1}^{N} \left(y_i - \boldsymbol{\beta}^{'}\boldsymbol{x}_i\right)^2 \\
\widehat{\boldsymbol{\beta}} &= \text{arg min}_{\boldsymbol{\beta}} \left\{\sum_{i=1}^{N} \left(y_i - \boldsymbol{\beta}^{'}\boldsymbol{x}_i\right)^2\right\} \\
&= \left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y}
\end{aligned}
$$

Regression models

Suppose we have N documents, with each document i having label $y_i \in \{-1, 1\} \rightsquigarrow \{\text{liberal, conservative}\}$
We represent each document i is $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$.

$$
\begin{aligned}
f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) &= \sum_{i=1}^{N} \left(y_i - \boldsymbol{\beta}^{'} \boldsymbol{x}_i\right)^2 \\
\widehat{\boldsymbol{\beta}} &= \text{arg min}_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} \left(y_i - \boldsymbol{\beta}^{'} \boldsymbol{x}_i\right)^2 \right\} \\
&= \left(\boldsymbol{X}^{'} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^{'} \boldsymbol{Y}
\end{aligned}
$$

Problem:

Regression models

Suppose we have N documents, with each document i having label $y_i \in \{-1, 1\} \rightsquigarrow \{\text{liberal, conservative}\}$
We represent each document i is $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$.

$$
\begin{aligned}
f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) &= \sum_{i=1}^{N} \left(y_i - \boldsymbol{\beta}^{'} \boldsymbol{x}_i \right)^2 \\
\widehat{\boldsymbol{\beta}} &= \text{arg min}_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} \left(y_i - \boldsymbol{\beta}^{'} \boldsymbol{x}_i \right)^2 \right\} \\
&= \left(\boldsymbol{X}^{'} \boldsymbol{X} \right)^{-1} \boldsymbol{X}^{'} \boldsymbol{Y}
\end{aligned}
$$

Problem:

- J will likely be large (perhaps $J > N$)

Regression models

Suppose we have N documents, with each document i having label $y_i \in \{-1, 1\} \rightsquigarrow \{\text{liberal, conservative}\}$
We represent each document i is $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$.

$$
\begin{aligned}
f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) &= \sum_{i=1}^{N} \left(y_i - \boldsymbol{\beta}' \boldsymbol{x}_i\right)^2 \\
\widehat{\boldsymbol{\beta}} &= \text{arg min}_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} \left(y_i - \boldsymbol{\beta}' \boldsymbol{x}_i\right)^2 \right\} \\
&= \left(\boldsymbol{X}' \boldsymbol{X}\right)^{-1} \boldsymbol{X}' \boldsymbol{Y}
\end{aligned}
$$

Problem:

- J will likely be large (perhaps $J > N$)
- There many correlated variables

Regression models

Suppose we have N documents, with each document i having label $y_i \in \{-1, 1\} \rightsquigarrow$ {liberal, conservative}
We represent each document i is $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$.

$$
\begin{aligned}
f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) &= \sum_{i=1}^{N} \left(y_i - \boldsymbol{\beta}^{'}\boldsymbol{x}_i\right)^2 \\
\widehat{\boldsymbol{\beta}} &= \text{arg min}_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} \left(y_i - \boldsymbol{\beta}^{'}\boldsymbol{x}_i\right)^2 \right\} \\
&= \left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^{'}\boldsymbol{Y}
\end{aligned}
$$

Problem:

- J will likely be large (perhaps $J > N$)
- There many correlated variables

Predictions will be variable

Mean Square Error

Suppose θ is some value of the true parameter

Mean Square Error

Suppose θ is some value of the true parameter
Bias:

Mean Square Error

Suppose θ is some value of the true parameter
Bias:

$$\text{Bias} = E[\widehat{\theta} - \theta]$$

Mean Square Error

Suppose θ is some value of the true parameter
Bias:

$$\text{Bias} = E[\widehat{\theta} - \theta]$$

We may care about average distance from truth

Mean Square Error

Suppose θ is some value of the true parameter
Bias:

$$\text{Bias} = E[\widehat{\theta} - \theta]$$

We may care about average distance from truth

$$\mathrm{E}[(\hat{\theta} - \theta)^2]$$

Mean Square Error

Suppose θ is some value of the true parameter
Bias:

$$\text{Bias} = E[\widehat{\theta} - \theta]$$

We may care about average distance from truth

$$\text{E}[(\hat{\theta} - \theta)^2] = E[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + \theta^2$$

Mean Square Error

Suppose θ is some value of the true parameter
Bias:

$$\text{Bias} = E[\widehat{\theta} - \theta]$$

We may care about average distance from truth

$$\begin{aligned} \text{E}[(\hat{\theta} - \theta)^2] &= E[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + \theta^2 \\ &= E[\hat{\theta}^2] - E[\hat{\theta}]^2 + E[\hat{\theta}]^2 - 2\theta E[\hat{\theta}] + \theta^2 \end{aligned}$$

Mean Square Error

Suppose θ is some value of the true parameter
Bias:

$$\text{Bias} = E[\widehat{\theta - \theta}]$$

We may care about average distance from truth

$$\begin{aligned}
\mathrm{E}[(\hat{\theta} - \theta)^2] &= E[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + \theta^2 \\
&= E[\hat{\theta}^2] - E[\hat{\theta}]^2 + E[\hat{\theta}]^2 - 2\theta E[\hat{\theta}] + \theta^2 \\
&= E[\hat{\theta}^2] - E[\hat{\theta}]^2 + (E[\widehat{\theta - \theta}])^2
\end{aligned}$$

Mean Square Error

Suppose θ is some value of the true parameter
Bias:

$$\text{Bias} = E[\widehat{\theta} - \theta]$$

We may care about average distance from truth

$$\begin{aligned}
\text{E}[(\hat{\theta} - \theta)^2] &= E[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + \theta^2 \\
&= E[\hat{\theta}^2] - E[\hat{\theta}]^2 + E[\hat{\theta}]^2 - 2\theta E[\hat{\theta}] + \theta^2 \\
&= E[\hat{\theta}^2] - E[\hat{\theta}]^2 + (E[\widehat{\theta} - \theta])^2 \\
&= \text{Var}(\theta) + \text{Bias}^2
\end{aligned}$$

Mean Square Error

Suppose θ is some value of the true parameter
Bias:

$$\text{Bias} = E[\widehat{\theta} - \theta]$$

We may care about average distance from truth

$$\begin{aligned}
\text{E}[(\hat{\theta} - \theta)^2] &= E[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + \theta^2 \\
&= E[\hat{\theta}^2] - E[\hat{\theta}]^2 + E[\hat{\theta}]^2 - 2\theta E[\hat{\theta}] + \theta^2 \\
&= E[\hat{\theta}^2] - E[\hat{\theta}]^2 + (E[\widehat{\theta} - \theta])^2 \\
&= \text{Var}(\theta) + \text{Bias}^2
\end{aligned}$$

To reduce MSE, we are willing to induce bias to decrease variance$\rightsquigarrow$ methods that shrink coefficeints toward zero

Ridge Regression

Penalty for model complexity

Ridge Regression

Penalty for model complexity

$$f(\beta, \boldsymbol{X}, \boldsymbol{Y})$$

Ridge Regression

Penalty for model complexity

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) \quad = \quad \sum_{i=1}^{N} \left(y_i - \ \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} \right)^2$$

Ridge Regression

Penalty for model complexity

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^{N} \left(y_i - \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} \right)^2 + \underbrace{\lambda \sum_{j=1}^{J} \beta_j^2}_{\text{Penalty}}$$

Ridge Regression

Penalty for model complexity

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) \;=\; \sum_{i=1}^{N} \left(y_i - \; \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} \right)^2 + \underbrace{\lambda \sum_{j=1}^{J} \beta_j^2}_{\text{Penalty}}$$

where:

Ridge Regression

Penalty for model complexity

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^{N} \left(y_i - \ \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} \right)^2 + \underbrace{\lambda \sum_{j=1}^{J} \beta_j^2}_{\text{Penalty}}$$

where:

- $\beta_0 \rightsquigarrow$ intercept

Ridge Regression

Penalty for model complexity

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) \;=\; \sum_{i=1}^{N} \left(y_i - \;\beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} \right)^2 + \underbrace{\lambda \sum_{j=1}^{J} \beta_j^2}_{\text{Penalty}}$$

where:

- $\beta_0 \rightsquigarrow$ intercept
- $\lambda \rightsquigarrow$ penalty parameter

Ridge Regression $\rightsquigarrow$ Optimization

$$\boldsymbol{\beta}^{\text{Ridge}} = \arg\min_{\boldsymbol{\beta}} \left\{ f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) \right\}$$

Ridge Regression$\rightsquigarrow$ Optimization

$$
\begin{aligned}
\boldsymbol{\beta}^{\text{Ridge}} &= \text{arg min}_{\boldsymbol{\beta}} \left\{ f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) \right\} \\
&= \text{arg min}_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{J} \beta_j^2 \right\}
\end{aligned}
$$

Ridge Regression$\leadsto$ Optimization

$$\begin{aligned}
\boldsymbol{\beta}^{\text{Ridge}} &= \text{arg min}_{\boldsymbol{\beta}} \{f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y})\} \\
&= \text{arg min}_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{J} \beta_j^2 \right\} \\
&= \text{arg min}_{\boldsymbol{\beta}} \left\{ (\boldsymbol{Y} - \boldsymbol{X}^{'}\boldsymbol{\beta})^{'}(\boldsymbol{Y} - \boldsymbol{X}^{'}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^{'}\boldsymbol{\beta} \right\}
\end{aligned}$$

Demean the data and set $\beta_0 = \bar{y} = \sum_{i=1}^{N} \frac{y_i}{N}$

Ridge Regression$\leadsto$ Optimization

$$
\begin{aligned}
\boldsymbol{\beta}^{\text{Ridge}} &= \text{arg min}_{\boldsymbol{\beta}} \left\{ f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) \right\} \\
&= \text{arg min}_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{J} \beta_j^2 \right\} \\
&= \text{arg min}_{\boldsymbol{\beta}} \left\{ (\boldsymbol{Y} - \boldsymbol{X}^{'}\boldsymbol{\beta})^{'}(\boldsymbol{Y} - \boldsymbol{X}^{'}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^{'}\boldsymbol{\beta} \right\} \\
&= \left(\boldsymbol{X}^{'}\boldsymbol{X} + \lambda \boldsymbol{I}_J \right)^{-1} \boldsymbol{X}^{'}\boldsymbol{Y}
\end{aligned}
$$

Demean the data and set $\beta_0 = \bar{y} = \sum_{i=1}^{N} \frac{y_i}{N}$

Ridge Regression$\rightsquigarrow$ Optimization

$$
\begin{aligned}
\boldsymbol{\beta}^{\text{Ridge}} &= \text{arg min}_{\boldsymbol{\beta}} \left\{ f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) \right\} \\
&= \text{arg min}_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{J} \beta_j^2 \right\} \\
&= \text{arg min}_{\boldsymbol{\beta}} \left\{ (\boldsymbol{Y} - \boldsymbol{X}^{'} \boldsymbol{\beta})^{'} (\boldsymbol{Y} - \boldsymbol{X}^{'} \boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^{'} \boldsymbol{\beta} \right\} \\
&= \left(\boldsymbol{X}^{'} \boldsymbol{X} + \lambda \boldsymbol{I}_J \right)^{-1} \boldsymbol{X}^{'} \boldsymbol{Y}
\end{aligned}
$$

Demean the data and set $\beta_0 = \bar{y} = \sum_{i=1}^{N} \frac{y_i}{N}$

Ridge Regression $\rightsquigarrow$ Intuition (1)

Suppose $\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{I}_J$.

Ridge Regression $\rightsquigarrow$ Intuition (1)

Suppose $\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{I}_J$.

$$\widehat{\beta} \;=\; \left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y}$$

Ridge Regression$\rightsquigarrow$ Intuition (1)

Suppose $\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{I}_J$.

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\ &= \boldsymbol{X}^{'}\boldsymbol{Y} \end{aligned}$$

Ridge Regression⇝ Intuition (1)

Suppose $\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{I}_J$.

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\
&= \boldsymbol{X}^{'}\boldsymbol{Y} \\
\boldsymbol{\beta}^{\text{ridge}} &= \left(\boldsymbol{X}^{'}\boldsymbol{X} + \lambda\boldsymbol{I}_J\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y}
\end{aligned}
$$

Ridge Regression$\rightsquigarrow$ Intuition (1)

Suppose $\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{I}_J$.

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\
&= \boldsymbol{X}^{'}\boldsymbol{Y} \\
\boldsymbol{\beta}^{\text{ridge}} &= \left(\boldsymbol{X}^{'}\boldsymbol{X} + \lambda\boldsymbol{I}_J\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\
&= \left(\boldsymbol{I}_j + \lambda\boldsymbol{I}_j\right)\boldsymbol{X}^{'}\boldsymbol{Y}
\end{aligned}
$$

Ridge Regression$\rightsquigarrow$ Intuition (1)

Suppose $\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{I}_J$.

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\
&= \boldsymbol{X}^{'}\boldsymbol{Y} \\
\boldsymbol{\beta}^{\text{ridge}} &= \left(\boldsymbol{X}^{'}\boldsymbol{X} + \lambda\boldsymbol{I}_J\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\
&= \left(\boldsymbol{I}_j + \lambda\boldsymbol{I}_j\right)\boldsymbol{X}^{'}\boldsymbol{Y} \\
&= \left(\boldsymbol{I}_j + \lambda\boldsymbol{I}_j\right)\widehat{\boldsymbol{\beta}}
\end{aligned}
$$

Ridge Regression$\rightsquigarrow$ Intuition (1)

Suppose $\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{I}_J$.

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\
&= \boldsymbol{X}^{'}\boldsymbol{Y} \\
\boldsymbol{\beta}^{\text{ridge}} &= \left(\boldsymbol{X}^{'}\boldsymbol{X} + \lambda\boldsymbol{I}_J\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\
&= \left(\boldsymbol{I}_j + \lambda\boldsymbol{I}_j\right)\boldsymbol{X}^{'}\boldsymbol{Y} \\
&= \left(\boldsymbol{I}_j + \lambda\boldsymbol{I}_j\right)\widehat{\boldsymbol{\beta}} \\
\beta_j^{\text{Ridge}} &= \frac{\widehat{\beta}_j}{1+\lambda}
\end{aligned}$$

Ridge Regression$\rightsquigarrow$ Intuition (2)

$$
\begin{aligned}
\boldsymbol{\beta}_j &\sim \text{Normal}(0, \tau^2) \\
y_i &\sim \text{Normal}(\beta_0 + \boldsymbol{x}_i^{'}\boldsymbol{\beta}, \sigma^2)
\end{aligned}
$$

Ridge Regression$\rightsquigarrow$ Intuition (2)

$$\begin{aligned}
\boldsymbol{\beta}_j &\sim \text{Normal}(0, \tau^2) \\
y_i &\sim \text{Normal}(\beta_0 + \boldsymbol{x}_i^{'}\boldsymbol{\beta}, \sigma^2)
\end{aligned}$$

$$p(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{Y}) \propto \prod_{j=1}^{J} p(\beta_j) \prod_{i=1}^{N} p(y_i|\boldsymbol{x}_i, \boldsymbol{\beta})$$

Ridge Regression$\rightsquigarrow$ Intuition (2)

$$
\begin{aligned}
\boldsymbol{\beta}_j & \sim \text{Normal}(0, \tau^2) \\
y_i & \sim \text{Normal}(\beta_0 + \boldsymbol{x}_i^{'}\boldsymbol{\beta}, \sigma^2)
\end{aligned}
$$

$$
\begin{aligned}
p(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{Y}) & \propto \prod_{j=1}^{J} p(\beta_j) \prod_{i=1}^{N} p(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}) \\
& \propto \prod_{j=1}^{J} \frac{1}{\sqrt{2\pi}\tau} \exp\left(-\frac{\beta_j^2}{2\tau^2}\right) \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \beta_0 + \boldsymbol{x}^{'}\boldsymbol{\beta})^2}{2\sigma^2}\right)
\end{aligned}
$$

Ridge Regression $\leadsto$ Intuition (2)

$$\log p(\beta|\boldsymbol{X}, \boldsymbol{Y}) = -\sum_{j=1}^{J} \frac{\beta_j^2}{2\tau^2} - \sum_{i=1}^{N} \frac{(y_i - \beta_0 + \boldsymbol{x}'\boldsymbol{\beta})^2}{2\sigma^2}$$

Ridge Regression$\rightsquigarrow$ Intuition (2)

$$\begin{aligned}
\log p(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{Y}) &= -\sum_{j=1}^{J} \frac{\beta_j^2}{2\tau^2} - \sum_{i=1}^{N} \frac{(y_i - \beta_0 + \boldsymbol{x}^{'} \boldsymbol{\beta})^2}{2\sigma^2} \\
-2\sigma^2 \log p(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{Y}) &= \sum_{i=1}^{N} (y_i - \beta_0 + \boldsymbol{x}^{'} \boldsymbol{\beta})^2 + \sum_{j=1}^{J} \frac{\sigma^2}{\tau^2} \beta_j^2
\end{aligned}$$

Ridge Regression$\leadsto$ Intuition (2)

$$\begin{aligned}\log p(\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{Y}) &= -\sum_{j=1}^{J}\frac{\beta_j^2}{2\tau^2} - \sum_{i=1}^{N}\frac{(y_i - \beta_0 + \boldsymbol{x}^{'}\boldsymbol{\beta})^2}{2\sigma^2} \\ -2\sigma^2\log p(\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{Y}) &= \sum_{i=1}^{N}(y_i - \beta_0 + \boldsymbol{x}^{'}\boldsymbol{\beta})^2 + \sum_{j=1}^{J}\frac{\sigma^2}{\tau^2}\beta_j^2\end{aligned}$$

where:

Ridge Regression$\rightsquigarrow$ Intuition (2)

$$\log p(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{Y}) = -\sum_{j=1}^{J} \frac{\beta_j^2}{2\tau^2} - \sum_{i=1}^{N} \frac{(y_i - \beta_0 + \boldsymbol{x}^{'}\boldsymbol{\beta})^2}{2\sigma^2}$$

$$-2\sigma^2 \log p(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^{N}(y_i - \beta_0 + \boldsymbol{x}^{'}\boldsymbol{\beta})^2 + \sum_{j=1}^{J} \frac{\sigma^2}{\tau^2}\beta_j^2$$

where:

- $\lambda = \frac{\sigma^2}{\tau^2}\beta_j^2$

Lasso Regression Objective Function/Optimization

Different Penalty for Model Complexity

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^{N} \left(y_i - \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{J} \underbrace{|\beta_j|}_{\text{Penalty}}$$

Lasso Regression Objective Function/Optimization

Different Penalty for Model Complexity

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^{N} \left(y_i - \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{J} \underbrace{|\beta_j|}_{\text{Penalty}}$$

- Optimization is non-linear (Absolute Value)

Lasso Regression Objective Function/Optimization

Different Penalty for Model Complexity

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^{N} \left(y_i - \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{J} \underbrace{|\beta_j|}_{\text{Penalty}}$$

- Optimization is non-linear (Absolute Value)
  - Coordinate Descent

Lasso Regression Objective Function/Optimization

Different Penalty for Model Complexity

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^{N} \left(y_i - \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{J} \underbrace{|\beta_j|}_{\text{Penalty}}$$

- Optimization is non-linear (Absolute Value)
  - Coordinate Descent
  - Start with Ridge

Lasso Regression Objective Function/Optimization

Different Penalty for Model Complexity

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^{N} \left(y_i - \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{J} \underbrace{|\beta_j|}_{\text{Penalty}}$$

- Optimization is non-linear (Absolute Value)
  - Coordinate Descent
  - Start with Ridge
  - Sub-differential, update steps

Lasso Regression Objective Function/Optimization

Different Penalty for Model Complexity

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) \;=\; \sum_{i=1}^{N}\left(y_i - \beta_0 + \sum_{j=1}^{J}\beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{J}\underbrace{|\beta_j|}_{\text{Penalty}}$$

- Optimization is non-linear (Absolute Value)
  - Coordinate Descent
  - Start with Ridge
  - Sub-differential, update steps
- Induces sparsity$\rightsquigarrow$ sets some coefficients to zero

Lasso Regression $\rightsquigarrow$ Intuition 1, Soft Thresholding

Suppose again $\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{I}_J$

Lasso Regression $\rightsquigarrow$ Intuition 1, Soft Thresholding

Suppose again $\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{I}_J$

$$f(\beta, \boldsymbol{X}, \boldsymbol{Y}) = (Y - \boldsymbol{X}\beta)^{'}(Y - \boldsymbol{X}\beta) + \lambda \sum_{j=1}^{J} |\beta_j|$$

Lasso Regression $\rightsquigarrow$ Intuition 1, Soft Thresholding

Suppose again $\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{I}_J$

$$\begin{aligned} f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) &= (Y - \boldsymbol{X}\boldsymbol{\beta})^{'}(Y - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{J} |\beta_j| \\ &= -2\boldsymbol{X}^{'}\boldsymbol{Y}\boldsymbol{\beta} + \boldsymbol{\beta}^{'}\boldsymbol{\beta} + \lambda \sum_{j=1}^{J} |\beta_j| \end{aligned}$$

Lasso Regression$\rightsquigarrow$ Intuition 1, Soft Thresholding

Suppose again $\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{I}_J$

$$
\begin{aligned}
f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) &= (Y - \boldsymbol{X}\boldsymbol{\beta})^{'}(Y - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{J} |\beta_j| \\
&= -2\boldsymbol{X}^{'}\boldsymbol{Y}\boldsymbol{\beta} + \boldsymbol{\beta}^{'}\boldsymbol{\beta} + \lambda \sum_{j=1}^{J} |\beta_j|
\end{aligned}
$$

The coefficient is

Lasso Regression $\rightsquigarrow$ Intuition 1, Soft Thresholding

Suppose again $\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{I}_J$

$$
\begin{aligned}
f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) &= (Y - \boldsymbol{X}\boldsymbol{\beta})^{'}(Y - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{J} |\beta_j| \\
&= -2\boldsymbol{X}^{'}\boldsymbol{Y}\boldsymbol{\beta} + \boldsymbol{\beta}^{'}\boldsymbol{\beta} + \lambda \sum_{j=1}^{J} |\beta_j|
\end{aligned}
$$

The coefficient is

$$
\beta_j^{\text{LASSO}} = \text{sign}\left(\widehat{\beta}_j\right)\left(|\widehat{\beta}_j| - \lambda\right)_{+}
$$

Lasso Regression $\rightsquigarrow$ Intuition 1, Soft Thresholding

Suppose again $\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{I}_J$

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) = (Y - \boldsymbol{X}\boldsymbol{\beta})^{'}(Y - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{J} |\beta_j|$$

$$= -2\boldsymbol{X}^{'}\boldsymbol{Y}\boldsymbol{\beta} + \boldsymbol{\beta}^{'}\boldsymbol{\beta} + \lambda \sum_{j=1}^{J} |\beta_j|$$

The coefficient is

$$\beta_j^{\text{LASSO}} = \text{sign}\left(\widehat{\beta}_j\right)\left(|\widehat{\beta}_j| - \lambda\right)_{+}$$

- $\text{sign}(\cdot) \rightsquigarrow 1$ or -1

Lasso Regression $\rightsquigarrow$ Intuition 1, Soft Thresholding

Suppose again $\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{I}_J$

$$
\begin{aligned}
f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) &= \left(Y - \boldsymbol{X}\boldsymbol{\beta}\right)^{'}\left(Y - \boldsymbol{X}\boldsymbol{\beta}\right) + \lambda \sum_{j=1}^{J} |\beta_j| \\
&= -2\boldsymbol{X}^{'}\boldsymbol{Y}\boldsymbol{\beta} + \boldsymbol{\beta}^{'}\boldsymbol{\beta} + \lambda \sum_{j=1}^{J} |\beta_j|
\end{aligned}
$$

The coefficient is

$$
\beta_j^{\text{LASSO}} = \text{sign}\left(\widehat{\beta}_j\right)\left(|\widehat{\beta}_j| - \lambda\right)_{+}
$$

- $\text{sign}(\cdot) \rightsquigarrow 1$ or -1
- $\left(|\widehat{\beta}_j| - \lambda\right)_{+} = \max(|\widehat{\beta}_j| - \lambda, 0)$

Lasso Regression$\rightsquigarrow$ Intuition 1, Soft Thresholding

Compare soft assignment

Lasso Regression $\rightsquigarrow$ Intuition 1, Soft Thresholding

Compare soft assignment

$$\beta_j^{\text{LASSO}} \quad = \quad \text{sign}\left(\widehat{\beta}_j\right)\left(|\widehat{\beta}_j| - \lambda\right)_+$$

Lasso Regression $\rightsquigarrow$ Intuition 1, Soft Thresholding

Compare soft assignment

$$\beta_j^{\text{LASSO}} = \text{sign}\left(\widehat{\beta}_j\right)\left(|\widehat{\beta}_j| - \lambda\right)_+$$

With hard assignment, selecting M biggest components

Lasso Regression$\rightsquigarrow$ Intuition 1, Soft Thresholding

Compare soft assignment

$$\beta_j^{\text{LASSO}} = \text{sign}\left(\widehat{\beta}_j\right)\left(|\widehat{\beta}_j| - \lambda\right)_+$$

With hard assignment, selecting M biggest components

$$\beta_j^{\text{subset}} = \widehat{\beta}_j \cdot I\left(|\widehat{\beta}_j| \geq |\widehat{\beta}_{(M)}|\right)$$

Lasso Regression $\rightsquigarrow$ Intuition 1, Soft Thresholding

Compare soft assignment

$$\beta_j^{\text{LASSO}} = \text{sign}\left(\widehat{\beta}_j\right)\left(|\widehat{\beta}_j| - \lambda\right)_+$$

With hard assignment, selecting M biggest components

$$\beta_j^{\text{subset}} = \widehat{\beta}_j \cdot I\left(|\widehat{\beta}_j| \geq |\widehat{\beta}_{(M)}|\right)$$

Intuition 2: Prior on coefficients $\rightsquigarrow$ Double exponential

Lasso Regression$\rightsquigarrow$ Intuition 1, Soft Thresholding

Compare soft assignment

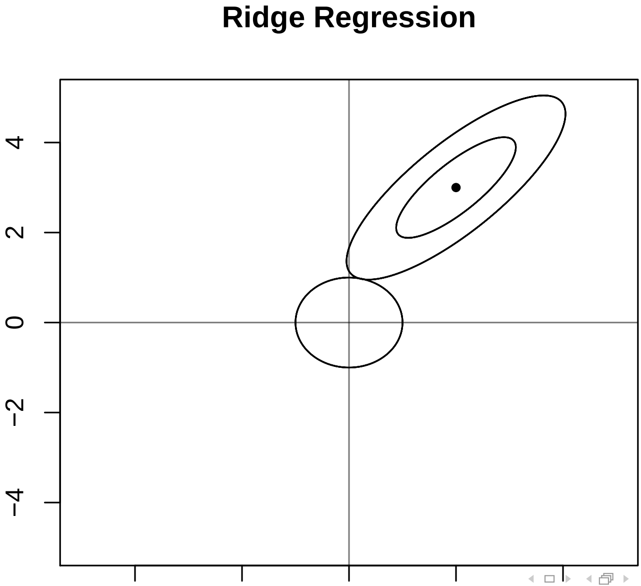
$$\beta_j^{\text{LASSO}} = \text{sign}\left(\widehat{\beta}_j\right)\left(|\widehat{\beta}_j| - \lambda\right)_+$$

With hard assignment, selecting M biggest components

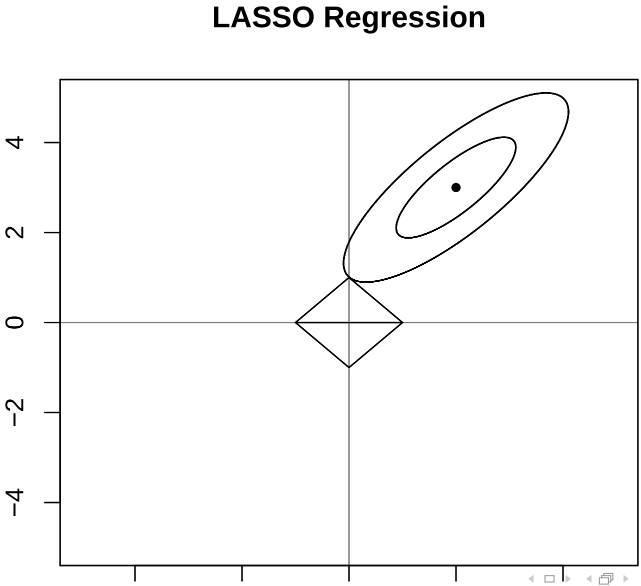
$$\beta_j^{\text{subset}} = \widehat{\beta}_j \cdot I\left(|\widehat{\beta}_j| \geq |\widehat{\beta}_{(M)}|\right)$$

Intuition 2: Prior on coefficients $\rightsquigarrow$ Double exponential
Why does LASSO induce sparsity?

Comparing Ridge and LASSO



Comparing Ridge and LASSO



Comparing Ridge and LASSO

Contrast $\beta = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $\tilde{\beta} = (1, 0)$

Comparing Ridge and LASSO

Contrast $\beta = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $\tilde{\beta} = (1, 0)$

Under ridge:

Comparing Ridge and LASSO

Contrast $\beta = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $\tilde{\beta} = (1, 0)$

Under ridge:

$$\sum_{j=1}^{2} \beta_j^2 \quad = \quad \frac{1}{2} + \frac{1}{2} = 1$$

Comparing Ridge and LASSO

Contrast $\beta = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $\tilde{\beta} = (1, 0)$

Under ridge:

$$\sum_{j=1}^{2} \beta_j^2 = \frac{1}{2} + \frac{1}{2} = 1$$

$$\sum_{j=1}^{2} \tilde{\beta}_j^2 = 1 + 0 = 1$$

Comparing Ridge and LASSO

Contrast $\beta = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $\tilde{\beta} = (1, 0)$

Under ridge:

$$\begin{aligned} \sum_{j=1}^{2} \beta_j^2 &= \frac{1}{2} + \frac{1}{2} = 1 \\ \sum_{j=1}^{2} \tilde{\beta}_j^2 &= 1 + 0 = 1 \end{aligned}$$

Under LASSO

Comparing Ridge and LASSO

Contrast $\beta = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $\tilde{\beta} = (1, 0)$

Under ridge:

$$\begin{aligned}
\sum_{j=1}^{2} \beta_j^2 &= \frac{1}{2} + \frac{1}{2} = 1 \\
\sum_{j=1}^{2} \tilde{\beta}_j^2 &= 1 + 0 = 1
\end{aligned}$$

Under LASSO

$$\sum_{j=1}^{2} |\beta_j| = \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} = \sqrt{2}$$

Comparing Ridge and LASSO

Contrast $\beta = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $\tilde{\beta} = (1, 0)$

Under ridge:

$$\begin{aligned} \sum_{j=1}^{2} \beta_j^2 &= \frac{1}{2} + \frac{1}{2} = 1 \\ \sum_{j=1}^{2} \tilde{\beta}_j^2 &= 1 + 0 = 1 \end{aligned}$$

Under LASSO

$$\begin{aligned} \sum_{j=1}^{2} |\beta_j| &= \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} = \sqrt{2} \\ \sum_{j=1}^{2} |\tilde{\beta}_j| &= 1 + 0 = 1 \end{aligned}$$

Selecting λ

How do we determine λ? $\rightsquigarrow$ Cross validation (lecture on Thursday)

`To the R code!`

Selecting λ

How do we determine λ? $\rightsquigarrow$ Cross validation (lecture on Thursday)
Applying models gives score (probability) of document belong to class$\rightsquigarrow$ threshold to classify

`To the R code!`

Selecting λ

How do we determine λ? $\rightsquigarrow$ Cross validation (lecture on Thursday)

Applying models gives score (probability) of document belong to class $\rightsquigarrow$ threshold to classify

`To the R code!`

Assessing Models (Elements of Statistical Learning)

- Model Selection: tuning parameters to select final model (next week's discussion)
- Model assessment : after selecting model, estimating error in classification

Comparing Training and Validation Set

Text classification and model assessment

- Replicate classification exercise with validation set
- General principle of classification/prediction
- Compare supervised learning labels to hand labels

Confusion matrix

Comparing Training and Validation Set

Representation of Test Statistics from Dictionary week (along with some new ones)

	Actual Label	
Classification (algorithm)	Liberal	Conservative
Liberal	True Liberal	False Liberal
Conservative	False Conservative	True Conservative

Comparing Training and Validation Set

Representation of Test Statistics from Dictionary week (along with some new ones)

	Actual Label	
Classification (algorithm)	Liberal	Conservative
Liberal	True Liberal	False Liberal
Conservative	False Conservative	True Conservative

$$\text{Accuracy} = \frac{\text{TrueLib} + \text{TrueCons}}{\text{TrueLib} + \text{TrueCons} + \text{FalseLib} + \text{FalseCons}}$$

Comparing Training and Validation Set

Representation of Test Statistics from Dictionary week (along with some new ones)

	Actual Label	
Classification (algorithm)	Liberal	Conservative
Liberal	True Liberal	False Liberal
Conservative	False Conservative	True Conservative

$$\text{Accuracy} = \frac{\text{TrueLib} + \text{TrueCons}}{\text{TrueLib} + \text{TrueCons} + \text{FalseLib} + \text{FalseCons}}$$

$$\text{Precision}_{\text{Liberal}} = \frac{\text{True Liberal}}{\text{True Liberal} + \text{False Liberal}}$$

Comparing Training and Validation Set

Representation of Test Statistics from Dictionary week (along with some new ones)

	Actual Label	
Classification (algorithm)	Liberal	Conservative
Liberal	True Liberal	False Liberal
Conservative	False Conservative	True Conservative

$$\text{Accuracy} = \frac{\text{TrueLib} + \text{TrueCons}}{\text{TrueLib} + \text{TrueCons} + \text{FalseLib} + \text{FalseCons}}$$

$$\text{Precision}_{\text{Liberal}} = \frac{\text{True Liberal}}{\text{True Liberal} + \text{False Liberal}}$$

$$\text{Recall}_{\text{Liberal}} = \frac{\text{True Liberal}}{\text{True Liberal} + \text{False Conservative}}$$

Comparing Training and Validation Set

Representation of Test Statistics from Dictionary week (along with some new ones)

	Actual Label	
Classification (algorithm)	Liberal	Conservative
Liberal	True Liberal	False Liberal
Conservative	False Conservative	True Conservative

$$
\begin{aligned}
\text{Accuracy} &= \frac{\text{TrueLib} + \text{TrueCons}}{\text{TrueLib} + \text{TrueCons} + \text{FalseLib} + \text{FalseCons}} \\
\text{Precision}_{\text{Liberal}} &= \frac{\text{True Liberal}}{\text{True Liberal} + \text{False Liberal}} \\
\text{Recall}_{\text{Liberal}} &= \frac{\text{True Liberal}}{\text{True Liberal} + \text{False Conservative}} \\
F_{\text{Liberal}} &= \frac{2\text{Precision}_{\text{Liberal}}\text{Recall}_{\text{Liberal}}}{\text{Precision}_{\text{Liberal}} + \text{Recall}_{\text{Liberal}}}
\end{aligned}
$$

Comparing Training and Validation Set

Representation of Test Statistics from Dictionary week (along with some new ones)

	Actual Label	
Classification (algorithm)	Liberal	Conservative
Liberal	True Liberal	False Liberal
Conservative	False Conservative	True Conservative

$$\text{Accuracy} = \frac{\text{TrueLib} + \text{TrueCons}}{\text{TrueLib} + \text{TrueCons} + \text{FalseLib} + \text{FalseCons}}$$

$$\text{Precision}_{\text{Liberal}} = \frac{\text{True Liberal}}{\text{True Liberal} + \text{False Liberal}}$$

$$\text{Recall}_{\text{Liberal}} = \frac{\text{True Liberal}}{\text{True Liberal} + \text{False Conservative}}$$

$$F_{\text{Liberal}} = \frac{2\text{Precision}_{\text{Liberal}}\text{Recall}_{\text{Liberal}}}{\text{Precision}_{\text{Liberal}} + \text{Recall}_{\text{Liberal}}}$$

ROC Curve

ROC as a measure of model performance

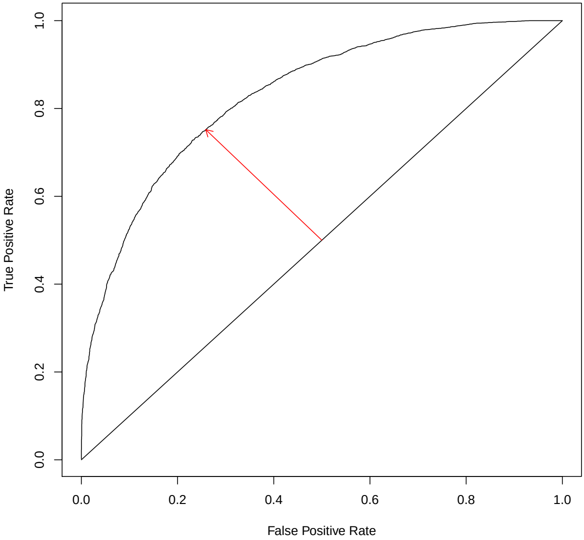
$$\text{Recall}_{\text{Liberal}} = \frac{\text{True Liberal}}{\text{True Liberal} + \text{False Conservative}}$$
$$\text{Recall}_{\text{Conservative}} = \frac{\text{True Conservative}}{\text{True Conservative} + \text{False Liberal}}$$

Tension:

- Everything liberal: $\text{Recall}_{\text{Liberal}} = 1$; $\text{Recall}_{\text{Conservative}} = 0$
- Everything conservative: $\text{Recall}_{\text{Liberal}} = 0$; $\text{Recall}_{\text{Conservative}} = 1$

Characterize Tradeoff:
Plot True Positive Rate $\text{Recall}_{\text{Liberal}}$
False Positive Rate ($1 - \text{Recall}_{\text{Conservative}}$)

Precision/Recall Tradeoff



Simple Classification Example

Analyzing house press releases
Hand Code: 1,000 press releases

- Advertising
- Credit Claiming
- Position Taking

Divide 1,000 press releases into two sets

- 500: Training set
- 500: Test set

Initial exploration: provides baseline measurement at classifier performances
Improve: through improving model fit

Example from First Model Fit

	Actual Label		
Classification (Naive Bayes)	Position Taking	Advertising	Credit Claim.
Position Taking	10	0	0
Advertising	2	40	2
Credit Claiming	80	60	306

$$
\begin{aligned}
\text{Accuracy} &= \frac{10+40+306}{500} = 0.71 \\
\text{Precision}_{PT} &= \frac{10}{10} = 1 \\
\text{Recall}_{PT} &= \frac{10}{10+2+80} = 0.11 \\
\text{Precision}_{AD} &= \frac{40}{40+2+2} = 0.91 \\
\text{Recall}_{AD} &= \frac{40}{40+60} = 0.4 \\
\text{Precision}_{Credit} &= \frac{306}{306+80+60} = 0.67 \\
\text{Recall}_{Credit} &= \frac{306}{306+2} = 0.99
\end{aligned}
$$

Fit Statistics in R

`RWeka` library provides **Amazing** functionality.
You can easily code them yourself

Support Vector Machines

Document i is an $J \times 1$ vector of counts

$$\boldsymbol{x}_i = (x_{1i}, x_{2i}, \ldots, x_{Ji})$$

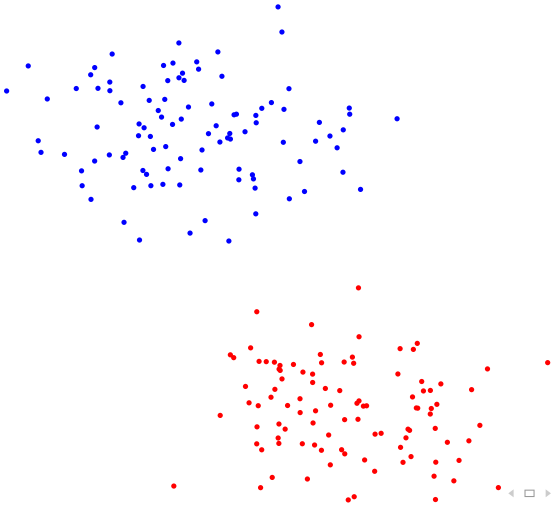
Suppose we have two classes, C_1, C_2.

$$\begin{aligned} Y_i &= 1 \text{ if } i \text{ is in class } 1 \\ Y_i &= -1 \text{ if } i \text{ is in class } 2 \end{aligned}$$

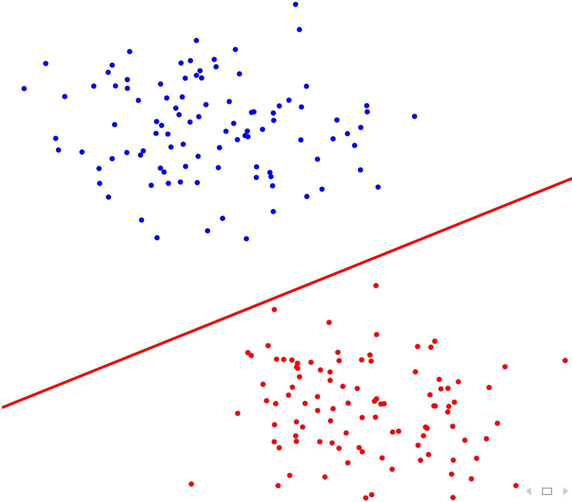
Suppose they are separable:

- Draw a line between groups
- Goal: identify the line in the middle
- Maximum margin

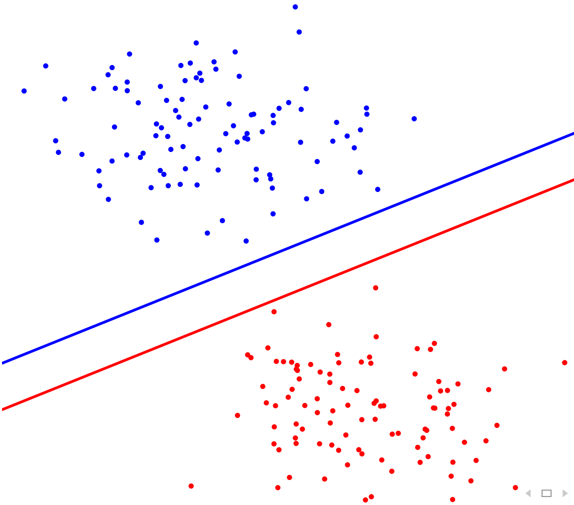
Support Vector Machines: Maximum Margin Classifier (Bishop 2006)



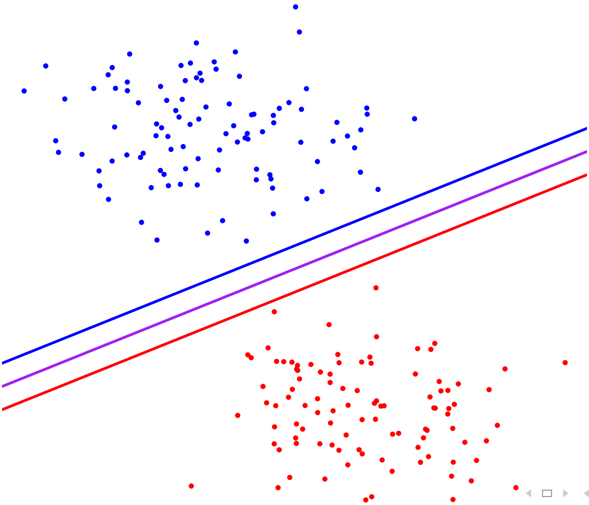
Support Vector Machines: Maximum Margin Classifier (Bishop 2006)



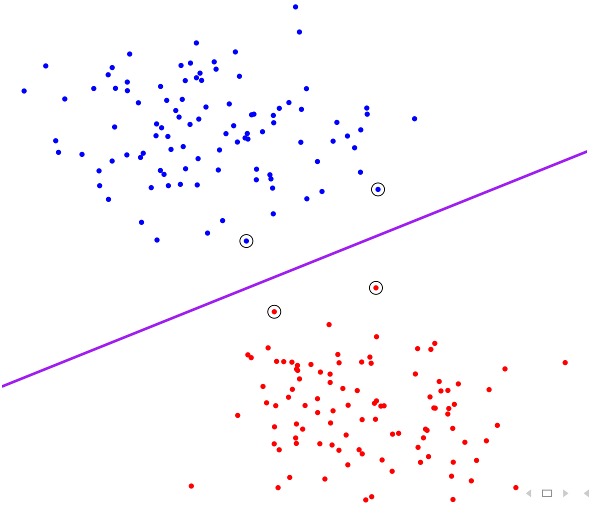
Support Vector Machines: Maximum Margin Classifier (Bishop 2006)



Support Vector Machines: Maximum Margin Classifier (Bishop 2006)



Support Vector Machines: Maximum Margin Classifier (Bishop 2006)



Support Vector Machines: Algebra (Bishop 2006)

Goal create a score to classify:

$$s(\boldsymbol{x}_i) \quad = \quad \boldsymbol{\beta}^{'}\boldsymbol{x}_i + b$$

- $\boldsymbol{\beta}$ Determines orientation of surface (slope)
- b determines location (moves surface up or down)
- If $s(\boldsymbol{x}_i) > 0 \rightarrow$ class 1
- If $s(\boldsymbol{x}_i) < 0 \rightarrow$ class 2
- $\frac{|s(\boldsymbol{x}_i)|}{||\boldsymbol{\beta}||} =$ Document distance from decision surface (margin)

Support Vector Machines: Algebra (Bishop 2006)

Objective function: maximum margin

Support Vector Machines: Algebra (Bishop 2006)

Objective function: maximum margin
$\min_i[\ |(s(\boldsymbol{x}_i)|\ ]$: Point closest to decision surface

Support Vector Machines: Algebra (Bishop 2006)

Objective function: maximum margin

$\min_i[\ |(s(\boldsymbol{x}_i)|\ ]$: Point closest to decision surface

We want to identify $\boldsymbol{\beta}$ and b to maximize the margin:

Support Vector Machines: Algebra (Bishop 2006)

Objective function: maximum margin

$\min_i[\ |(s(\boldsymbol{x}_i)|\ ]$: Point closest to decision surface

We want to identify $\boldsymbol{\beta}$ and b to maximize the margin:

$$\arg\max_{\boldsymbol{\beta},b} \left\{ \frac{1}{||\boldsymbol{\beta}||} \min_i[\ |(s(\boldsymbol{x}_i)|\ ] \right\}$$

Support Vector Machines: Algebra (Bishop 2006)

Objective function: maximum margin

$\min_i[\ |(s(\boldsymbol{x}_i)|\ ]$: Point closest to decision surface

We want to identify $\boldsymbol{\beta}$ and b to maximize the margin:

$$\text{arg max}_{\boldsymbol{\beta},b}\left\{\frac{1}{||\boldsymbol{\beta}||}\min_i[\ |(s(\boldsymbol{x}_i)|\ ]\right\}$$

$$\text{arg max}_{\boldsymbol{\beta},b}\left\{\frac{1}{||\boldsymbol{\beta}||}\min_i[\ |\boldsymbol{\beta}^{'}\boldsymbol{x}_i + b|\ ]\right\}$$

Support Vector Machines: Algebra (Bishop 2006)

Objective function: maximum margin

$\min_i[\ |(s(\boldsymbol{x}_i)|\ ]$: Point closest to decision surface

We want to identify $\boldsymbol{\beta}$ and b to maximize the margin:

$$\arg\max_{\boldsymbol{\beta},b}\left\{\frac{1}{||\boldsymbol{\beta}||}\min_i[\ |(s(\boldsymbol{x}_i)|\ ]\right\}$$

$$\arg\max_{\boldsymbol{\beta},b}\left\{\frac{1}{||\boldsymbol{\beta}||}\min_i[\ |\boldsymbol{\beta}^{'}\boldsymbol{x}_i + b|\ ]\right\}$$

Constrained optimization problem $\rightsquigarrow$ Quadratic programming problem

What About Overlap? (Bishop 2006)

- Rare that classes are separable.

What About Overlap? (Bishop 2006)

- Rare that classes are separable.
- Define:

What About Overlap? (Bishop 2006)

- Rare that classes are separable.
- Define:

$$\xi_i \quad = \quad 0 \text{ if correctly classified}$$

What About Overlap? (Bishop 2006)

- Rare that classes are separable.
- Define:

$$\begin{aligned}\xi_i &= 0 \text{ if correctly classified} \\ \xi_i &= |s(\boldsymbol{x}_i)| \text{ if incorrectly classified}\end{aligned}$$

What About Overlap? (Bishop 2006)

- Rare that classes are separable.
- Define:

$$\xi_i = 0 \text{ if correctly classified}$$
$$\xi_i = |s(\boldsymbol{x}_i)| \text{ if incorrectly classified}$$

Tradeoff:

What About Overlap? (Bishop 2006)

- Rare that classes are separable.
- Define:

$$
\begin{aligned}
\xi_i &= 0 \text{ if correctly classified} \\
\xi_i &= |s(\boldsymbol{x}_i)| \text{ if incorrectly classified}
\end{aligned}
$$

Tradeoff:

- Maximize margin between correctly classified groups

What About Overlap? (Bishop 2006)

- Rare that classes are separable.
- Define:

$$\begin{aligned}\xi_i &= 0 \text{ if correctly classified}\\ \xi_i &= |s(\boldsymbol{x}_i)| \text{ if incorrectly classified}\end{aligned}$$

Tradeoff:

- Maximize margin between correctly classified groups
- Minimize error from misclassified documents

What About Overlap? (Bishop 2006)

- Rare that classes are separable.
- Define:

$$
\begin{aligned}
\xi_i &= 0 \text{ if correctly classified} \\
\xi_i &= |s(\boldsymbol{x}_i)| \text{ if incorrectly classified}
\end{aligned}
$$

Tradeoff:

- Maximize margin between correctly classified groups
- Minimize error from misclassified documents

$$
\text{arg max}_{\boldsymbol{\beta}, b} \left\{ C \sum_{i=1}^{N} \xi_i + \frac{1}{||\boldsymbol{\beta}||} \text{ min}_i [\; |\boldsymbol{\beta}^{'} \boldsymbol{x}_i + b| \;] \right\}
$$

What About Overlap? (Bishop 2006)

- Rare that classes are separable.
- Define:

$$
\begin{aligned}
\xi_i &= 0 \text{ if correctly classified} \\
\xi_i &= |s(\boldsymbol{x}_i)| \text{ if incorrectly classified}
\end{aligned}
$$

Tradeoff:

- Maximize margin between correctly classified groups
- Minimize error from misclassified documents

$$
\text{arg max}_{\boldsymbol{\beta},b} \left\{ C \sum_{i=1}^{N} \xi_i + \frac{1}{||\boldsymbol{\beta}||} \min_i [\ |\boldsymbol{\beta}^{'} \boldsymbol{x}_i + b|\ ] \right\}
$$

C captures tradeoff

What About Overlap? (Bishop 2006)

- Rare that classes are separable.
- Define:

$$\begin{aligned} \xi_i &= 0 \text{ if correctly classified} \\ \xi_i &= |s(\boldsymbol{x}_i)| \text{ if incorrectly classified} \end{aligned}$$

Tradeoff:

- Maximize margin between correctly classified groups
- Minimize error from misclassified documents

$$\text{arg max}_{\boldsymbol{\beta},b}\left\{ C\sum_{i=1}^{N}\xi_i + \frac{1}{||\boldsymbol{\beta}||}\ \text{min}_i[\ |\boldsymbol{\beta}^{'}\boldsymbol{x}_i + b|\ ]\right\}$$

C captures tradeoff

How to Handle Multiple Comparisons?

- Rare that we only want to classify two categories

How to Handle Multiple Comparisons?

- Rare that we only want to classify two categories
- How to handle classification into K groups?

How to Handle Multiple Comparisons?

- Rare that we only want to classify two categories
- How to handle classification into K groups?
    1) Set up K classification problems:

How to Handle Multiple Comparisons?

- Rare that we only want to classify two categories
- How to handle classification into K groups?
  1) Set up K classification problems:
     - Compare each class to all other classes

How to Handle Multiple Comparisons?

- Rare that we only want to classify two categories
- How to handle classification into K groups?
  1) Set up K classification problems:
     - Compare each class to all other classes
     - Problem: can lead to inconsistent results

How to Handle Multiple Comparisons?

- Rare that we only want to classify two categories
- How to handle classification into K groups?
    1) Set up K classification problems:
        - Compare each class to all other classes
        - Problem: can lead to inconsistent results
        - Solution(?): select category with largest "score"

How to Handle Multiple Comparisons?

- Rare that we only want to classify two categories
- How to handle classification into K groups?
    1) Set up K classification problems:
        - Compare each class to all other classes
        - Problem: can lead to inconsistent results
        - Solution(?): select category with largest "score"
        - Problem: scales are not comparable

How to Handle Multiple Comparisons?

- Rare that we only want to classify two categories
- How to handle classification into K groups?
  1) Set up K classification problems:
     - Compare each class to all other classes
     - Problem: can lead to inconsistent results
     - Solution(?): select category with largest "score"
     - Problem: scales are not comparable
  2) Common solution: set up $K(K-1)/2$ classifications

How to Handle Multiple Comparisons?

- Rare that we only want to classify two categories
- How to handle classification into K groups?
  1) Set up K classification problems:
     - Compare each class to all other classes
     - Problem: can lead to inconsistent results
     - Solution(?): select category with largest "score"
     - Problem: scales are not comparable
  2) Common solution: set up $K(K-1)/2$ classifications
  - Perform vote to select class (still suboptimal)

How to Handle Multiple Comparisons?

- Rare that we only want to classify two categories
- How to handle classification into K groups?
  1) Set up K classification problems:
     - Compare each class to all other classes
     - Problem: can lead to inconsistent results
     - Solution(?): select category with largest "score"
     - Problem: scales are not comparable
  2) Common solution: set up $K(K-1)/2$ classifications
  - Perform vote to select class (still suboptimal)
  3) Simultaneous estimation possible, much slower

R Code to Run SVMs

```
library(e1071)
fit<- svm(T . , as.data.frame(tdm) , method ='C',
kernel='linear')
```

where: `method = 'C'` → Classification

`kernel='linear'` → allows for distortion of feature space. Options:

- Linear
- Polynomial
- Radial
- sigmoid

```
preds<- predict(fit, data =
as.data.frame(tdm[-c(1:no.train),]))
```

Example of SVMs in Political Science Research

Hillard, Purpura, Wilkerson: SVMs to code topic/sub topics for policy agendas project

TABLE 3. Bill Title Interannotator Agreement for Five Model Types

	SVM	MaxEnt	Boostexter	Naïve Bayes
Major topic *N* = 20	88.7% (.881)	86.5% (.859)	85.6% (.849)	81.4% (.805)
Subtopic *N* = 226	81.0% (.800)	78.3% (.771)	73.6% (.722)	71.9% (.705)

SVMs are under utilized in political science

ReadMe: Optimization for a Different Goal (Hopkins and King 2010)

Naive Bayes (and next week, SVM): focused on individual document classification.

ReadMe: Optimization for a Different Goal (Hopkins and King 2010)

Naive Bayes (and next week, SVM): focused on individual document classification.
But what if we're focused on proportions only?

ReadMe: Optimization for a Different Goal (Hopkins and King 2010)

Naive Bayes (and next week, SVM): focused on individual document classification.

But what if we're focused on proportions only?

Hopkins and King (2010): method for characterizing distribution of classes

ReadMe: Optimization for a Different Goal (Hopkins and King 2010)

Naive Bayes (and next week, SVM): focused on individual document classification.

But what if we're focused on proportions only?

Hopkins and King (2010): method for characterizing distribution of classes

Can be much more accurate than individual classifiers, requires fewer assumptions (do not need random sample of documents) .

ReadMe: Optimization for a Different Goal (Hopkins and King 2010)

Naive Bayes (and next week, SVM): focused on individual document classification.
But what if we're focused on proportions only?
Hopkins and King (2010): method for characterizing distribution of classes
Can be much more accurate than individual classifiers, requires fewer assumptions (do not need random sample of documents) .

- King and Lu (2008): derive method for characterizing causes of deaths for verbal autopsies

ReadMe: Optimization for a Different Goal (Hopkins and King 2010)

Naive Bayes (and next week, SVM): focused on individual document classification.

But what if we're focused on proportions only?

Hopkins and King (2010): method for characterizing distribution of classes

Can be much more accurate than individual classifiers, requires fewer assumptions (do not need random sample of documents) .

- King and Lu (2008): derive method for characterizing causes of deaths for verbal autopsies
- Hopkins and King (2010): extend the method to text documents

ReadMe: Optimization for a Different Goal (Hopkins and King 2010)

Naive Bayes (and next week, SVM): focused on individual document classification.
But what if we're focused on proportions only?
Hopkins and King (2010): method for characterizing distribution of classes
Can be much more accurate than individual classifiers, requires fewer assumptions (do not need random sample of documents) .

- King and Lu (2008): derive method for characterizing causes of deaths for verbal autopsies
- Hopkins and King (2010): extend the method to text documents

Basic intuition:

ReadMe: Optimization for a Different Goal (Hopkins and King 2010)

Naive Bayes (and next week, SVM): focused on individual document classification.
But what if we're focused on proportions only?
Hopkins and King (2010): method for characterizing distribution of classes
Can be much more accurate than individual classifiers, requires fewer assumptions (do not need random sample of documents) .

- King and Lu (2008): derive method for characterizing causes of deaths for verbal autopsies
- Hopkins and King (2010): extend the method to text documents

Basic intuition:

- Examine joint distribution of characteristics (without making Naive Bayes like assumption)
- Focus on distributions (only) makes this analysis possible

ReadMe: Optimization for a Different Goal (Hopkins and King 2010)

Measure only presence/absence of each term [(Jx1) vector]

ReadMe: Optimization for a Different Goal (Hopkins and King 2010)

Measure only presence/absence of each term [$(J x 1)$ vector]

$$\boldsymbol{x}_i = (1, 0, 0, 1, \ldots, 0)$$

ReadMe: Optimization for a Different Goal (Hopkins and King 2010)

Measure only presence/absence of each term [(Jx1) vector]

$$\boldsymbol{x}_i = (1, 0, 0, 1, \ldots, 0)$$

What are the possible realizations of $\boldsymbol{x}_i$?

ReadMe: Optimization for a Different Goal (Hopkins and King 2010)

Measure only presence/absence of each term [(Jx1) vector]

$$\boldsymbol{x}_i = (1, 0, 0, 1, \ldots, 0)$$

What are the possible realizations of $\boldsymbol{x}_i$?

- 2^J possible vectors

ReadMe: Optimization for a Different Goal (Hopkins and King 2010)

Measure only presence/absence of each term [(Jx1) vector]

$$\boldsymbol{x}_i = (1, 0, 0, 1, \ldots, 0)$$

What are the possible realizations of $\boldsymbol{x}_i$?

- 2^J possible vectors

Define:

ReadMe: Optimization for a Different Goal (Hopkins and King 2010)

Measure only presence/absence of each term [(Jx1) vector]

$$\boldsymbol{x}_i = (1, 0, 0, 1, \ldots, 0)$$

What are the possible realizations of $\boldsymbol{x}_i$?

- 2^J possible vectors

Define:

$$P(\boldsymbol{x}) = \text{probability of observing } \boldsymbol{x}$$

ReadMe: Optimization for a Different Goal (Hopkins and King 2010)

Measure only presence/absence of each term [(Jx1) vector]

$$\boldsymbol{x}_i = (1, 0, 0, 1, \ldots, 0)$$

What are the possible realizations of $\boldsymbol{x}_i$?

- 2^J possible vectors

Define:

$$\begin{aligned} P(\boldsymbol{x}) &= \text{probability of observing } \boldsymbol{x} \\ P(\boldsymbol{x}|C_j) &= \text{Probability of observing } \boldsymbol{x} \text{ conditional on category } C_j \end{aligned}$$

ReadMe: Optimization for a Different Goal (Hopkins and King 2010)

Measure only presence/absence of each term [(Jx1) vector]

$$\boldsymbol{x}_i = (1, 0, 0, 1, \ldots, 0)$$

What are the possible realizations of $\boldsymbol{x}_i$?

- 2^J possible vectors

Define:

$$\begin{aligned}
P(\boldsymbol{x}) &= \text{probability of observing } \boldsymbol{x} \\
P(\boldsymbol{x}|C_j) &= \text{Probability of observing } \boldsymbol{x} \text{ conditional on category } C_j \\
P(\boldsymbol{X}|C) &= \text{Matrix collecting vectors}
\end{aligned}$$

ReadMe: Optimization for a Different Goal (Hopkins and King 2010)

Measure only presence/absence of each term [(Jx1) vector]

$$\boldsymbol{x}_i = (1, 0, 0, 1, \ldots, 0)$$

What are the possible realizations of $\boldsymbol{x}_i$?

- 2^J possible vectors

Define:

$$\begin{aligned} P(\boldsymbol{x}) &= \text{probability of observing } \boldsymbol{x} \\ P(\boldsymbol{x}|C_j) &= \text{Probability of observing } \boldsymbol{x} \text{ conditional on category } C_j \\ P(\boldsymbol{X}|C) &= \text{Matrix collecting vectors} \\ P(C) &= P(C_1, C_2, \ldots, C_K) \text{ target quantity of interest} \end{aligned}$$

ReadMe: Optimization for a Different Goal (Hopkins and King 2010)

Measure only presence/absence of each term [(Jx1) vector]

$$\boldsymbol{x}_i = (1, 0, 0, 1, \ldots, 0)$$

What are the possible realizations of $\boldsymbol{x}_i$?

- 2^J possible vectors

Define:

$$\begin{aligned}
P(\boldsymbol{x}) &= \text{probability of observing } \boldsymbol{x} \\
P(\boldsymbol{x}|C_j) &= \text{Probability of observing } \boldsymbol{x} \text{ conditional on category } C_j \\
P(\boldsymbol{X}|C) &= \text{Matrix collecting vectors} \\
P(C) &= P(C_1, C_2, \ldots, C_K) \text{ target quantity of interest}
\end{aligned}$$

ReadMe: Optimization for a Different Goal (Hopkins and King 2010)

$$\underbrace{P(\boldsymbol{x})}_{2^J x 1} = \underbrace{P(\boldsymbol{x}|C)}_{2^J x K} \underbrace{P(C)}_{K x 1}$$

Matrix algebra problem to solve, for $P(C)$
Like Naive Bayes, requires two pieces to estimate
Complication $2^J >> $ no. documents
Kernel Smoothing Methods (without a formal model)

- $P(\boldsymbol{x})$ = estimate directly from test set
- $P(\boldsymbol{x}|C)$ = estimate from training set
  - Key assumption: $P(\boldsymbol{x}|C)$ in training set is equivalent to $P(\boldsymbol{x}|C)$ in test set
- If true, can perform biased sampling of documents, worry less about drift...

Algorithm Summarized

- Estimate $\hat{p}(\boldsymbol{x})$ from test set
- Estimate $\hat{p}(\boldsymbol{x}|C)$ from training set
- Use $\hat{p}(\boldsymbol{x})$ and $\hat{p}(\boldsymbol{x}|C)$ to solve for $p(C)$

Assessing Model Performance

Not classifying individual documents $\rightarrow$ different standards
Mean Square Error :

$$E[(\hat{\theta} - \theta)^2] = \text{var}(\hat{\theta}) + \text{Bias}(\hat{\theta}, \theta)^2$$

Suppose we have true proportions $P(C)^{\text{true}}$. Then, we'll estimate Root Mean Square Error

$$\text{RMSE} = \sqrt{\frac{\sum_{j=1}^{J}(P(C_j)^{\text{true}} - P(C_j))}{J}}$$

$$\text{Mean Abs. Prediction Error} = \left|\frac{\sum_{j=1}^{J}(P(C_j)^{\text{true}} - P(C_j))}{J}\right|$$

Visualize: plot true and estimated proportions

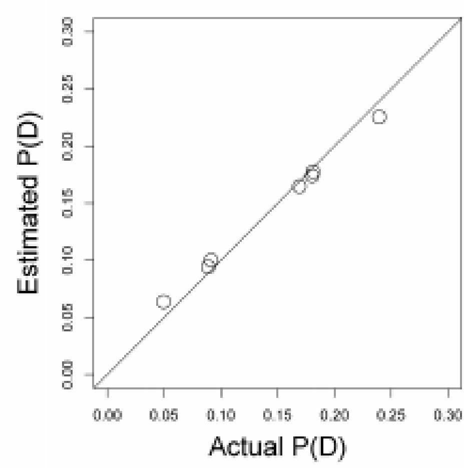

Estimated P(D)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Actual P(D)

Table 1 Performance of Our Nonparametric Approach and Four Support Vector Machine Analyses

	Percent of Blog Posts Correctly Classified			
	In-Sample Fit	In-Sample Cross-Validation	Out-of-Sample Prediction	Mean Absolute Proportion Error
Nonparametric	—	—	—	1.2
Linear	67.6	55.2	49.3	7.7
Radial	67.6	54.2	49.1	7.7
Polynomial	99.7	48.9	47.8	5.3
Sigmoid	15.6	15.6	18.2	23.2

Notes: Each row is the optimal choice over numerous individual runs given a specific kernel. Leaving aside the sigmoid kernel, individual classification performance in the first three columns does not correlate with mean absolute error in the document category proportions in the last column.

Using the House Press Release Data

Method	RMSE	APSE
ReadMe	0.036	0.056
NaiveBayes	0.096	0.14
SVM	0.052	0.084

Code to Run in R

Control file:

filename	truth	trainingset
20July2009LEWIS53.txt	4	1
26July2006LEWIS249.txt	2	0

```
tdm<- undergrad(control=control, fullfreq=F)
process<- preprocess(tdm)
output<- undergrad(process)
output$est.CSMF ## proportion in each category
output$true.CSMF ## if labeled for validation set (but not
used in training set)
```

Model Selection!