

MS&E 114/214 Autumn 2024 Homework 4

Handed out 11/7, due 11/15 at noon

Problem 1, 2(a), 3, and 4 are worth 10 points each. Problems 2(b), 5(a), and 5(b) are worth 5 points each.

- 1) *Training a classifier using Support Vector Machines.* This problem is entirely described in the following colab file
https://colab.research.google.com/drive/1vFs6ECjiSBGdY3PCG6G9y0_RQVzfyup?usp=sharing.
Solve it on colab and submit it to us by pasting the url on the uploaded pdf to gradescope.
- 2) *Iris classification using logistic regression.*

a) **Extra Credit for 114, Regular problem for 214:**

Consider the iris data from scikit-learn which consists of two different types (versicolor and setosa) of irises with four features for every data point. Let us denote every data point by $\mathbf{u}^i \in \mathbb{R}^4$ for every $i \in [N]$. We now consider a model with parameters $\mathbf{P} \in \mathbb{R}^4$ (corresponding to the coefficients for every feature) and constant term $\alpha \in \mathbb{R}$.

Recall that for pattern classification in class, we defined the penalty to be $\max(0, -\mathbf{P}^T \mathbf{u}^i + \alpha + 1)$ and $\max(0, \mathbf{P}^T \mathbf{u}^i - \alpha + 1)$ when the data-point \mathbf{u}^i belongs to versicolor and setosa respectively. In this problem we shall consider a loss function based on the logistic regression model.

The penalty for the case when iris data-point \mathbf{u}^i belongs to versicolor class is given by

$$L_1(\mathbf{z}^i) = -\log\left(\frac{1}{1 + e^{-z}}\right) \text{ where } \mathbf{z}^i = \mathbf{P}^T \mathbf{u}^i - \alpha \quad (1)$$

and when the iris-datapoint \mathbf{u}^i belongs to setosa class is given by

$$L_2(\mathbf{z}^i) = -\log\left(\frac{e^{-z}}{1 + e^{-z}}\right) \text{ where } \mathbf{z}^i = \mathbf{P}^T \mathbf{u}^i - \alpha \quad (2)$$

Plot the function $L_1(z)$ as a function of $z \in (-2, 2)$.

Now, train the model in excel or Python by solving an optimization problem which minimizes the sum of the penalties for every data point. Do not use Logistic Regression functionality in Excel or in any ML libraries.

Report the optimum value of the objective function and the learnt weight vectors, and also report the accuracy. Add a link to the excel or colab file in your pdf. Now argue why this loss function is “easy” to optimize.

Hint: Can you observe from the plot that the function $L_1(z)$ looks convex?

- b) **Extra Credit for both 114 and 214:** Add a “sum-of-squares” regularization penalty to this problem and solve for different values of the regularization parameter.
- 3) *More context about some optimization techniques.*
 - a) Watch this video on gradient descent and write a short summary. You need to only watch the part from 5:20 to 8:00, but you are of course welcome to watch the parts before and after for more context. If you solved Problem 2, then connect the fact that gradient descent finds a local minimum to the fact that Logistic Regression can generally be solved easily using gradient descent.

AI Agents might generate a good answer to this problem without you watching the video. We want the initial summary of the video to be in your own words. You can refine it using AI Agents but in that case, you have to also include your initial unrefined answer.

- b) Transform the following LP into a standard form and then take its dual. Show your work (that is, clearly write down the transformed LP, label the primal constraints with dual variables, and then take the dual.).

$$\begin{aligned} \text{Minimize} \quad & 2x + 3y \\ \text{s.t.} \quad & x \geq 0, \\ & x + y \geq 1. \end{aligned}$$

The standard forms used in class have non-negativity constraints for all variables, and either a minimization objective with \geq constraints or a maximization objective with \leq constraints.

Hint: Note that this problem is missing the constraint $y \geq 0$, and you can handle this issue by replacing y with the difference of two new decision variables, both ≥ 0 . Also note that if a LP is in dual standard form, its dual will be in primal standard form and obtained using the same procedure that we discussed in class.

- 4) *Super-replication and Arbitrage.* Consider the call and put options for Apple Stock (AAPL) in this spreadsheet, all maturing on 12/13/2024. Assume that these are all European options, i.e., they can only be exercised on 12/13. You are also given the current price of the Apple Stock, as well as a zero-coupon bond which costs \$0.997 and is going to pay \$1 on 12/13.

For planning purposes, assume that the price of AAPL stock on 12/13 can be any multiple of \$2.50 between \$200 and \$275, inclusive.

- a) Write down the payoff matrix P such that given any initial portfolio x of the zero-coupon bonds, Apple stocks, and the call and put options that are available to you, Px represents the payoff of this portfolio on 12/13.
 - b) Does this problem have an arbitrage opportunity? Show your work.
 - c) Now assume that there is a transaction cost of \$0.001 for each unit of bond bought or sold short in your portfolio, and a transaction cost of \$0.5 for each unit of stock or option bought or sold short in your portfolio. There will be no transaction costs on 12/13/24, that is, you can still use the payoff matrix P from parts (a) and (b). Is there still an arbitrage opportunity? Comment on how transaction costs can help keep a market more stable.
 - d) Note that there is no put option available in this market with a strike price of \$245. Construct the cheapest portfolio that super-replicates the payoff of this put option.
- 5) a) **Modeling:** Consider the Participatory Budgeting problem from class. Can you extend the LP from class to the case where each voter has a non-negative weight and your goal is to maximize the total weighted utility?
- b) **Extra Credit for both 114 and 214:** Consider the modeling problem above. You are now given a desired target allocation z of funds to the M projects, and your goal is to figure out whether there is some assignment of non-negative weights to voters such that the weights sum to 1, and such that with these weights, the target allocation becomes the optimum solution to your LP from the modeling problem. Model this new problem as an LP.

Hint: This is a hard problem. You need to think of a weighted vote threshold T such that every expenditure level that receives support less than T is not funded, and every expenditure level that receives support more than T is funded. No help will be provided for this problem.