CIS 4190/5190: Applied Machine Learning — Fall 2022

Homework 2

Handed Out: September 14 — *Due: October 3, 8:00 p.m.*

- You are encouraged to format your solutions using LaTeX. Handwritten solutions are permitted, but remember that you bear the risk that we may not be able to read your work and grade it properly — we will not accept post hoc explanations for illegible work. You will submit your solution manuscript for written HW 2 as a single PDF file.
- The homework is **due at 8:00 PM** on the due date. We will be using Gradescope for collecting the homework assignments. Please submit your solution manuscript as a PDF file via Gradescope. Post on Ed Discussion and contact the TAs if you are having technical difficulties in submitting the assignment.

1 Multiple Choice & Written Questions

Note: You do not need to show work for multiple choice questions. If formatting your answer in LaTeX, use our LaTeX template `hw2_template.tex` (This is a read-only link. You'll need to make a copy before you can edit. Make sure you make only private copies.).

1. [Bias-Variance Tradeoff] (3pts) Suppose we have an L_2-regularized linear regression model, which has loss $L(\beta) = \frac{1}{n}\sum_{i=1}^{n}(f_\beta(x_i) - y_i)^2 + \lambda\|\beta\|_2^2$. For each of the following, indicate whether it tends to increase or decrease bias, and similarly for variance:

   A) Increase the number of training examples n

   B) Increase the regularization parameter λ

   C) Increase the dimension d of the features $\phi(x) \in \mathbb{R}^d$

   In addition, suppose you fit a model and find that it has low loss on the training data but high loss on the test data; for each of the above three values n, λ, and d, indicate whether you should increase or decrease it to reduce the test loss.

2. [Gradient Ascent] (2pts) A function $L(\beta)$ is *concave* if $F(\beta) := -L(\beta)$ is convex. *Gradient ascent* takes steps of the form $\beta \leftarrow \beta + \eta \cdot \nabla_\beta L(\beta)$ (whereas gradient descent takes steps $\beta \leftarrow \beta - \eta \cdot \nabla_\beta L(x)$). If $L(\beta)$ is concave, what does gradient ascent converge to? [Hint: Rewrite the gradient ascent formula in terms of $F(\beta)$; what does it look like?]

   A) global maximum

   B) global minimum

   C) local minimum

   D) local maximum

3. [Regularization/Sparsity] (4 pts) In class, we demonstrated the intuition behind L_1 and L_2 regularization through two dimensional ellipsoid visualizations. In this problem we will take a different angle, and try to see why L_1 regularization helps create sparsity from the perspective of gradient descent.

   A) (2pts) Write down the partial derivative of the L_1 and L_2 regularization terms with respect to some weight β_j (you can ignore the case $\beta_j = 0$ where the gradient may be undefined). [Recall that the L_1 regularization term is $R_1(\beta) = \lambda \sum_{j'=1}^{d} |\beta_{j'}|$, and the L_2 regularization term is $R_2(\beta) = \lambda \sum_{j'=1}^{d} \beta_{j'}^2$.]

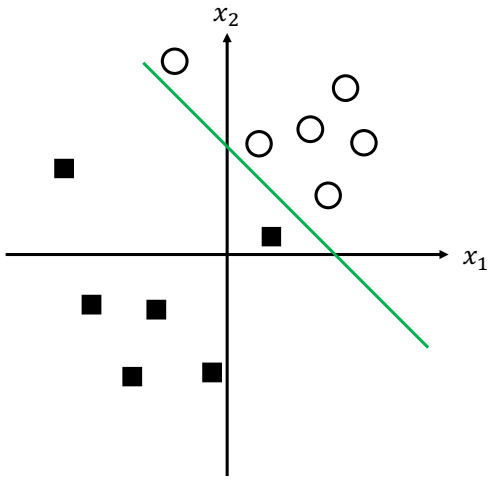
   B) (2pts) Based on these results, which regularizer does a better job "pushing" β_j to zero? [Hint: Consider what happens when β_j is already small. For simplicity, you can assume that the partial derivative $\frac{\partial}{\partial \beta_j}$ of the MSE term is zero.]

4. [Linear Regression] (4 pts) Suppose the true function we are trying to approximate is $y = \max\{x, 0\}$. Furthermore, suppose we use unregularized linear regression with the MSE loss function to fit a function $y = ax + b$. In the following scenarios, assume we are always sampling training inputs x uniformly at random from the given intervals, and assume we take $n \to \infty$, where n is the number of training examples.

   A) (2pts) Suppose we train on points sampled from $x \in [0, 1]$. Then, what is the learned model? Write out the coefficients a and b. What is the MSE on points sampled from $x \in [-1, 0]$?

   B) (2pts) Suppose we train on points sampled from $x \in [-1, 0]$. Then, what is the learned model? Write out the coefficients a and b What is the MSE on points sampled from $x \in [0, 1]$?

5. [Logistic Regression/Regularization] (4 pts) Suppose the input dimension is $d = 2$ (i.e., $x = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^\top \in \mathbb{R}^2$, and suppose we have the following true model and dataset:



   Here, the green line depicts the true model $y = \mathbb{1}(x_1 + x_2 \geq 1)$, the circles are labeled $y = 1$, and the solid squares are labeled $y = 0$. Recall that the accuracy is $\text{acc}(\beta; Z) =$

$\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\left(y_i = f_\beta(x_i)\right)$, where $f_\beta(x) = \mathbb{1}(\beta^\top x \geq 0)$. Note that if we train a logistic regression model without a feature map and with no regularization, then $\beta = \begin{bmatrix}1 & 1\end{bmatrix}^\top$ fits the training data with a single error, so its accuracy is $\frac{11}{12}$. By inspection, it is impossible to do better, so the best possible accuracy we can achieve is $\frac{11}{12}$.

A) (1 pt) Suppose we train a logistic regression model with an intercept feature map $\phi(x) = \begin{bmatrix}x_1 & x_2 & 1\end{bmatrix}$ (in this parameterization, β_3 is the intercept term). What is the best possible training accuracy of the model if we use no regularization?

B) (3 pts) Suppose we use the same intercept feature map as above, and furthermore, we use regularization

$$R(\beta) = \lambda_1\beta_1 + \lambda_2\beta_2 + \lambda_3\beta_3.$$

What is the best possible training accuracy if (i) $\lambda_1 = \lambda_2 = 0$ and $\lambda_3 \to \infty$; (ii) $\lambda_1 = \lambda_3 = 0$ and $\lambda_2 \to \infty$; and (iii) $\lambda_2 = \lambda_3 = 0$, and $\lambda_1 \to \infty$?

6. [Linear Regression; Mandatory for CIS 5190, Optional for CIS 4190] (6 pts) Recall that a closed form solution for the linear regression parameters is $\hat{\beta} = (X^\top X)^{-1}X^\top Y$, where $Z = (X, Y)$, and for a new example x, the predicted label is $f_{\hat{\beta}}(x) = \hat{\beta}^\top x$. Find a function $k_i(x; X)$ **depending only on** x**,** X**, and** i such that

$$f_{\hat{\beta}}(x) = \sum_{i=1}^{n} k_i(x; X)y_i.$$

In other words, the model $f_{\hat{\beta}}(x)$ can be expressed as a weighted combination of the training labels y_i.

2 Python Programming Questions

A IPython notebook is linked on the class website. It will tell you everything you need to do, and provide starter code. Remember to include the plots and answer the questions in your written homework submission!