CSM148 Homework 2 - Solutions Guide

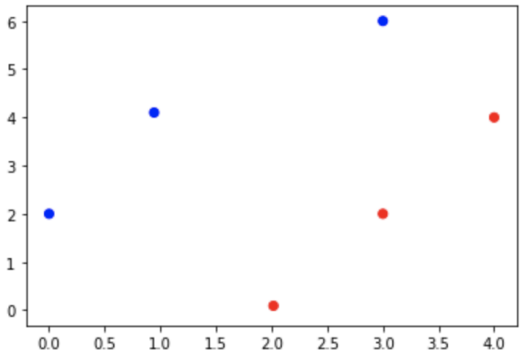
Due date: Thursday, May 6 at 10:00 AM PST
Instructions: All work must be completed individually. If you consulted with any classmates for the homework, please note them on the first page.

Start each problem on a new page, and be sure to clearly label where each problem and subproblem begins. All problems must be submitted in order (all of P1 before P2, etc.).

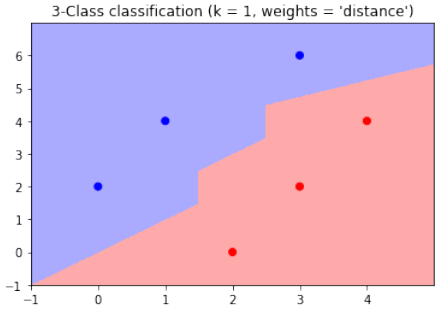
No late homeworks will be accepted. This is not out of a desire to be harsh, but rather out of fairness to all students in this large course.

1 K-Nearest Neighbors for Classification

Consider the following two types of data points shown in the Figure. The blue one with coordinates (x, y): (0, 2), (1,4), (3,6), and the red one with coordinates (x, y): (3, 2), (4, 4), (2, 0).

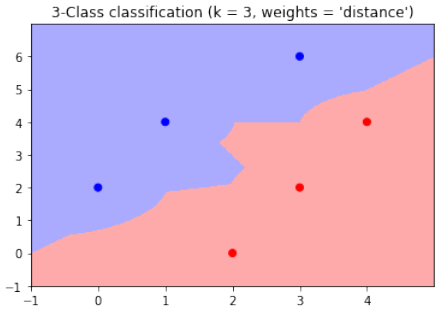


(a) Draw the decision boundary for 1-NN based on these data points.



(b) Draw the decision boundary for 3-NN based on these data points.

(c) If instead our dataset was comprised of blue points: (0, 200), (2, 500), (3,600) and red points: (3, 200), (4, 400), (1, 0), briefly explain what problem we might have with this kind of sampling and how can we address this problem? **Since KNN relies on relative distance, as these two features are**



not scaled equivalently the Y-axis will dominate any determination of relative distance. This can be addressed by scaling features

(d) Suppose you have a dataset consisting of 1000 samples. Describe a method to select the optimal K to use for KNN? **Iterate through different K values systematically and see what K value produces the optimal results**

2 True or False, Simple Explanations

Provide brief explanations for all your answers.

Recall that we have loss functions for LASSO and Ridge regression as $\sum_{i=1}^{n}(y_i - \boldsymbol{\beta}x_i)^2/n + \lambda\sum_{j=1}^{d}|\beta_j|$ and $\sum_{i=1}^{n}(y_i - \boldsymbol{\beta}x_i)^2/n + \lambda\sum_{j=1}^{d}\beta_j^2$, respectively.

(a) (T or F) LASSO regression will not tend to give you sparse coefficients $\boldsymbol{\beta}$. **False. ℓ_1 regularization will tend to make the solution vector $\boldsymbol{\beta}$ sparse.**

(b) (T or F) Ridge regression will tend to give you sparse coefficients $\boldsymbol{\beta}$. **False. ℓ_2 regularization will have effect on the magnitude of the solution vector $\boldsymbol{\beta}$ instead of the sparsity.**

(c) (T or F) When we have large λ in Ridge regression, we expect the magnitude of $\boldsymbol{\beta}$ to be small. **True. A large λ will contribute to a large penalty on the magnitude of the solution vector $\boldsymbol{\beta}$.**

(d) (T or F) When we increase λ in LASSO regression, we expect the number of nonzero coefficients in $\boldsymbol{\beta}$ to decrease. **True. With the increase of λ in LASSO regression, the number of nonzero coefficients in $\boldsymbol{\beta}$ will decrease.**

(e) Data scientists do almost all of their work in Python and are unlikely to use other tools. **False. While Python is a critical tool for datascience, datascientists will employ a wide variety of tools in the course of their work.**

(f) Data Engineering is a minor component of the datascience pipeline. **False. Engineering data is often among the most critical steps in the data science process.**

3 Logistic Regression Boundary and Interpretation

Suppose we fit a mulitple logistic regression $\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$.

(a) Suppose we have p features, write down the general expression of X_p in terms of other features on the decision boundary.

$X_p = -\frac{1}{\beta_p}(\beta_0 + \beta_1 X_1 + \ldots, \beta_{p-1}X_{p-1})$

(b) When $p = 2$, specifically what is the relationship between X_1 and X_2 on the decision boundary?

$X_2 = -\frac{1}{\beta_2}(\beta_0 + \beta_1 X_1)$.

(c) Suppose we have $p = 2$, and $\beta_0 = -1, \beta_1 = 1, \beta_2 = 2$, give interpretations for these coefficients. When $X_1 = X_2 = 0$, what are the odds and probability of the event that $Y = 1$? How does one unit increase in X_1 or X_2 change the odds and probability of the event that $Y = 1$?

When $X_1 = X_2 = 0$, we have $\log \frac{P(Y=1)}{1-P(Y=1)} = -1$ (This is also the interpretation of β_0). $P(Y = 1) = 1/(1 + e)$. One unit increase in X_1 will increase the odds by a factor of e^1.

4 Confusion Table

Suppose we have the following confusion table output by the logistic regression using the probability threshold $P(Y = 1) \geq \pi$.

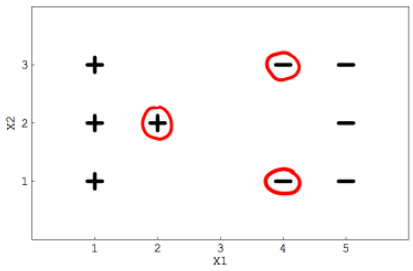
	Predicted $\hat{Y} = 0$	Predicted $\hat{Y} = 1$
Actual $Y = 0$	801	66
Actual $Y = 1$	150	45

(a) What are false positives, false negatives, true positives, and true negatives. **False positives: 66, False negatives: 150, True positives: 45, True negatives: 801**

(b) Compute true positive rate and false positive rate. **TPR = Recall = 45/195 = 0.23 FPR = 66/867 = 0.076**

(c) How will the confusion table change if we increase the threshold π? **The total number of Y predictors = 1 will decrease, reducing both false positives and true positives, however likely reducing false positives to a greater degree**

5 Support Vector Machine



(a) Suppose you have a dataset with 2 classes ('+' and '-'). If you remove one of the points that **is not** circled will that alter your Decision boundary? **It will not effect it as SVMs rely solely on the subset of points that define the decision boundary.**

(b) What is meant by a hard margin or soft margin? In this case will it matter if your Decision Boundary is either? **Hard and soft margins refer to the degree of flexibility the SVM classifier has to assign labels that may fall on the 'wrong' side of the decision boundary, particularly when there isn't a clear delineation of the dataset into classes, with soft-margining allowing for more flexible classification. In this case**

(c) Explain what the parameters gamma and C of the Radial Basis Function (RBF) kernel SVM do.

The C parameter trades off correct classification of training examples against maximization of the decision function's margin. For larger values of C, a smaller margin will be accepted if the decision function is better at classifying all training points correctly. A lower C will encourage a larger margin, therefore a simpler decision function, at the cost of training accuracy. In other words C behaves as

a regularization parameter in the SVM.**Intuitively, the gamma parameter defines how far the influence of a single training example reaches, with low values meaning 'far' and high values meaning 'close'. The gamma parameters can be seen as the inverse of the radius of influence of samples selected by the model as support vectors.**

The C parameter trades off correct classification of training examples against maximization of the decision function's margin. For larger values of C, a smaller margin will be accepted if the decision function is better at classifying all training points correctly. A lower C will encourage a larger margin, therefore a simpler decision function, at the cost of training accuracy. In other words C behaves as a regularization parameter in the SVM.

6 Augmentation

Many methods for making predictions from data, such as linear regression, are limited in terms of the transformations that they can apply to input data before making a prediction. As linear regression assumes that the output is the sum of coefficients multiplied by input features, it is unable to account for cases where the impact of two features together is greater than the sum of their parts. For example, a house that both has > 5 bedrooms and is in California may be worth four times more than would be expected from the learned price impact of each feature on its own.

Feature Crosses are synthetic features you can form by crossing two or more features together, and they can help to improve the predictive power of techniques such as linear regression. Expanding on the above housing example, you could generate a new feature that indicates a combination of both a home's number of bedrooms and location.

(a) Describe two pairs of features from Project 1 that might be interesting to cross together, and explain why. **No specific guidance here, just a reasonable discussion about crossing two features**

(b) You have latitude and longitude for homes, and you think feature crosses may allow you to make better predictions. However, your latitude and longitude are continuously valued. How might you do a feature cross in this case? **Obvious approach is to discretize the features by blocking values ranges**