CS/ECE/ME532 Period 19 Activity

Estimated Time: 15 mins for P1, 20 mins for P2, 20 mins for P3

1. You have two feature vectors $\boldsymbol{x}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $\boldsymbol{x}_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$ and corresponding labels $d_1 = -1, d_2 = 1$. The linear classifier is $\text{sign}\{\boldsymbol{x}_i^T \boldsymbol{w}\}$ where $\boldsymbol{w} = \begin{bmatrix} -1 \\ 0.5 \end{bmatrix}$.

a) Find the squared error loss for this classifier.

b) Find the hinge loss for this classifier.

SOLUTION:

a) The squared-error loss is $\sum_{i=1}^{2}(d_i - \boldsymbol{x}_i^T \boldsymbol{w})^2 = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$

b) The hinge loss is $\sum_{i=1}^{2}(1 - d_i \boldsymbol{x}_i^T \boldsymbol{w})_+ = \frac{1}{2} + 0 = \frac{1}{2}$

2. You have four data points $x_1 = 2, x_2 = 1.5, x_3 = 1/2, x_4 = -1/2$ and corresponding labels $y_1 = 1, y_2 = 1, y_3 = -1, y_4 = -1$.

a) Find a maximum margin linear classifier for this data. *Hint:* Graph the data.

b) Use squared-error loss to train the classifier (with the help of Python). Does this classifier make any errors?

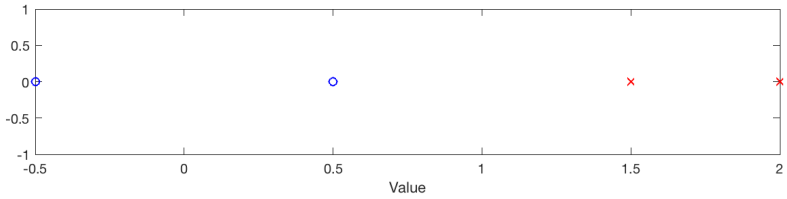
c) Find a classifier with zero hinge loss. *Hint:* Use what you've learned about hinge loss, not computation. Does this classifier make any errors?

d) Now suppose $x_4 = -5$. Use squared-error loss to find the classifier (with the help of Python). Does this classifier make any errors?

e) Can you still find a classifier with zero hinge loss when $x_4 = -5$? Does it make any errors?

SOLUTION:

a) Clearly $\text{sign}(x_i - a)$ is able to perfectly classify the data for $0.5 < a < 1.5$. The maximum margin occurs when the decision boundary is the midpoint of the two classes, or $a = 1$. The classifier needs to be offset from 0, which means two weights.



b) Define $\boldsymbol{A} = \begin{bmatrix} 2 & 1 \\ 1.5 & 1 \\ 0.5 & 1 \\ -0.5 & 1 \end{bmatrix}$, $\boldsymbol{d} = \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix}$ and $\boldsymbol{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$ gives a classifier with decision boundary $w_1 x_i + w_2 = 0$. We can solve for $\boldsymbol{w} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{d}$ using MATLAB or Python to obtain $\boldsymbol{w} = \begin{bmatrix} 0.9492 \\ -0.8305 \end{bmatrix}$ or a decision boundary of $x_i = 0.875$.
This achieves perfect classification.

c) Note the hinge loss is zero if $1 - d_i \boldsymbol{x}_i^T \boldsymbol{w} \leq 0$ or $d_i \boldsymbol{x}_i^T \boldsymbol{w} \geq 1$, which implies $|w_1 x_i + w_2| \geq 1$. This occurs for all data points for $w_1 = 2, w_2 = -2$. This also achieves perfect classification.

d) In this case we have $\boldsymbol{w} = \begin{bmatrix} 0.256 \\ 0.064 \end{bmatrix}$ or a decision boundary of $x_i = -0.25$. This misclassifies x_3.

e) Changing a point away from the boundary does not affect the hinge loss. The classifier $\boldsymbol{w} = \begin{bmatrix} 2 \\ -2 \end{bmatrix}$ derived above has zero hinge loss and achieves perfect classification.

3. Previously, we examined the performance of classifiers trained using the squared error loss function (i.e, trained using least squares). This problem uses an *off-the-shelf* Linear Support Vector Machine to train a binary linear classifier.

The data set is divided into training and test data sets. In order to represent a decision boundary that may not pass through the origin, we can consider the feature vector $\boldsymbol{x}^T = \begin{bmatrix} x_1 & x_2 & 1 \end{bmatrix}$.

a) *Classifier using off the shelf SVM.* Code is provided to train a classifier using an off the shelf SVM with hinge loss. Run the code to find the linear classifier weights. Next, uses the weights to predict the class of the test data. How many classification errors occur?

b) Comment out the code that trains the classifier using the linear SVM, and uncomment the code that train the classifier using least squares (i.e, $\boldsymbol{w}_{opt} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$). How many errors occur on the test set?

c) Training a classifier using the *squared error as a loss function* can fail when correctly labeled data points lie far from the decision boundary. Linear SVMs trained with hinge loss are not susceptible to the same problem. A new dataset consisting of the first dataset, plus 1000 (correctly labeled) datapoints at $x_1 = 0, x_2 = 10$ is created. What happens to the decision boundary when these new data points are included in training the linear SVM?

d) How does this compare with the error rate of the linear classifier trained with the new data points? Why is the such a difference in performance?

SOLUTION:

a) There are 1213 errors using the wights from the off the shelf SVM.

b) There are 495 errors, so the least squares classifier performs better with this training data.

c) Using the SVM, there are still 1213 errors.

d) With the least squares classifier and the 'easy' to classify data points, there are 2668 errors. The decision boundary is impacted by the squared error associated with these easy to classify point.