CS6140: Machine Learning

Homework Assignment # 2

Assigned: 02/05/2020 *Due: 02/18/2020, 11:59pm, through Blackboard*

Three problems, 95 points in total. Good luck!
Prof. Predrag Radivojac, Northeastern University

Problem 1. (25 points) Naive Bayes classifier. Consider a binary classification problem where there are only four data points in the training set. That is $\mathcal{D} = \{(-1,-1,-),(-1,+1,+),(+1,-1,+),(+1,+1,-)\}$, where each tuple (x_1, x_2, y) represents a training example with input vector (x_1, x_2) and class label y.

a) (10 points) Construct a naive Bayes classifier for this problem and evaluate its accuracy on the training set. Consider "accuracy" to be the fraction of correct predictions.

Solution:

I: We should choose the max between these two:

$$P(Y = +|(-1,-1)) = 1/2 * 1/2 * 1/2 = 1/8$$

$$P(Y = -|(-1,-1)) = 1/2 * 1/2 * 1/2 = 1/8$$

So the result is: +

II: We should choose the max between these two:

$$P(Y = +|(-1,+1)) = 1/2 * 1/2 * 1/2 = 1/8$$

$$P(Y = -|(-1,+1)) = 1/2 * 1/2 * 1/2 = 1/8$$

So the result is: +

III: We should choose the max between these two:

$$P(Y = +|(+1,-1)) = 1/2 * 1/2 * 1/2 = 1/8$$

$$P(Y = -|(+1,-1)) = 1/2 * 1/2 * 1/2 = 1/8$$

So the result is: +

IV: We should choose the max between these two:

$$P(Y = +|(+1,+1)) = 1/2 * 1/2 * 1/2 = 1/8$$

$$P(Y = -|(+1,+1)) = 1/2 * 1/2 * 1/2 = 1/8$$

So the result is: +

We will now calculate the accuracy. The formula had two correct results.

$$accuracy = 2/4 = 1/2$$

b) (10 points) Transform the input space into a six-dimensional space $(+1, x_1, x_2, x_1x_2, x_1^2, x_2^2)$ and repeat the previous step.

Solution:

So the new dataset is this:

$D = \{(+1,-1,-1,+1,+1,+1,-),(+1,-1,+1,-1,+1,+1,+),$

$(+1,+1,-1,-1,+1,+1,+),(+1,+1,+1,+1,+1,+1,-)\}$

I: We should choose the max between these two:

$$P(Y=+|(+1,-1,-1,+1,+1,+1)) = 2/2 * 1/2 * 1/2 * 0/2 = 0$$

$$P(Y=-|(+1,-1,-1,+1,+1,+1)) = 2/2 * 1/2 * 1/2 * 2/2 * 2/2 * 2/2 * 1/2 = 1/8$$

So the result is: -

II: We should choose the max between these two:

$$P(Y=+|(+1,-1,+1,-1,+1,+1)) = 2/2 * 1/2 * 1/2 * 2/2 * 2/2 * 2/2 * 1/2 = 1/8$$

$$P(Y=-|(+1,-1,+1,-1,+1,+1)) = 2/2 * 1/2 * 1/2 * 0/2 = 0$$

So the result is: +

III: We should choose the max between these two:

$$P(Y=+|(+1,+1,-1,-1,+1,+1)) = 2/2 * 1/2 * 1/2 * 2/2 * 2/2 * 2/2 * 1/2 = 1/8$$

$$P(Y=-|(+1,+1,-1,-1,+1,+1)) = 2/2 * 1/2 * 1/2 * 0/2 = 0$$

So the result is: +

IV: We should choose the max between these two:

$$P(Y=+|(+1,+1,+1,+1,+1,+1)) = 2/2 * 1/2 * 1/2 * 0/2 = 0$$

$$P(Y=-|(+1,+1,+1,+1,+1,+1)) = 2/2 * 1/2 * 1/2 * 2/2 * 2/2 * 2/2 * 1/2 = 1/8$$

So the result is: -

We will now calculate the accuracy. The formula had four correct results.

$$accuracy = 4/4 = 1$$

c) (5 points) Repeat the previous step when the data set accidentally includes the seventh feature, set to $-x_1x_2$. What is the impact of adding this dependent feature on the classification model?

Solution:

So the new dataset is this:

$D = \{(+1,-1,-1,+1,+1,+1,-1,-),(+1,-1,+1,-1,+1,+1,+1,+),$

$(+1,+1,-1,-1,+1,+1,+1,+),(+1,+1,+1,+1,+1,+1,-1,-)\}$

I: We should choose the max between these two:

$$P(Y=+|(+1,-1,-1,+1,+1,+1,-1)) = 2/2 * 1/2 * 1/2 * 0/2 = 0$$

$$P(Y=-|(+1,-1,-1,+1,+1,+1,-1)) = 2/2 * 1/2 * 1/2 * 2/2 * 2/2 * 2/2 * 0 * 1/2 = 0$$

So the result is: +

II: We should choose the max between these two:

$$P(Y = +|(+1, -1, +1, -1, +1, +1, +1)) = 2/2 * 1/2 * 1/2 * 2/2 * 2/2 * 2/2 * 0 * 1/2 = 0$$

$$P(Y = -|(+1, -1, +1, -1, +1, +1, +1)) = 2/2 * 1/2 * 1/2 * 0/2 = 0$$

So the result is: +

III: We should choose the max between these two:

$$P(Y = +|(+1, +1, -1, -1, +1, +1, +1)) = 2/2 * 1/2 * 1/2 * 2/2 * 2/2 * 2/2 * 0 * 1/2 = 0$$

$$P(Y = -|(+1, +1, -1, -1, +1, +1, +1)) = 2/2 * 1/2 * 1/2 * 0/2 = 0$$

So the result is: +

IV: We should choose the max between these two:

$$P(Y = +|(+1, +1, +1, +1, +1, +1, -1)) = 2/2 * 1/2 * 1/2 * 0/2 = 0$$

$$P(Y = -|(+1, +1, +1, +1, +1, +1, -1)) = 2/2 * 1/2 * 1/2 * 2/2 * 2/2 * 2/2 * 0 * 1/2 = 0$$

So the result is: +

We will now calculate the accuracy. The formula had two correct results. Adding this feature made our formula to generate wrong results for some inputs. It makes all the probabilities equal to zero so all of the predictions are now leading to + for the result.

$$accuracy = 2/4 = 1/2$$

Problem 2. (25 points) Consider a binary classification problem in which we want to determine the optimal decision surface. A point $\mathbf{x}$ is on the decision surface if $P(Y = 1|\mathbf{x}) = P(Y = 0|\mathbf{x})$.

a) (10 points) Find the optimal decision surface assuming that each class-conditional distribution is defined as a two-dimensional Gaussian distribution:

$$p(\mathbf{x}|Y = i) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}_i|^{1/2}} \cdot e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T\mathbf{\Sigma}_i^{-1}(\mathbf{x}-\mathbf{m}_i)}$$

where $i \in \{0, 1\}$, $\mathbf{m}_0 = (1, 2)$, $\mathbf{m}_1 = (6, 3)$, $\mathbf{\Sigma}_0 = \mathbf{\Sigma}_1 = \mathbf{I}_2$, $P(Y = 0) = P(Y = 1) = 1/2$, $\mathbf{I}_d$ is the d-dimensional identity matrix, and $|\mathbf{\Sigma}_i|$ is the determinant of $\mathbf{\Sigma}_i$.

b) (5 points) Generalize the solution from part (a) using $\mathbf{m}_0 = (m_{01}, m_{02})$, $\mathbf{m}_1 = (m_{11}, m_{12})$, $\mathbf{\Sigma}_0 = \mathbf{\Sigma}_1 = \sigma^2\mathbf{I}_2$ and $P(Y = 0) \neq P(Y = 1)$.

c) (10 points) Generalize the solution from part (b) to arbitrary covariance matrices $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_1$. Discuss the shape of the optimal decision surface.

Problem 3. (45 points) Consider a multivariate linear regression problem of mapping $\mathbb{R}^d$ to $\mathbb{R}$, with two different objective functions. The first objective function is the sum of squared errors, as presented in class; i.e., $\sum_{i=1}^{n} e_i^2$, where $e_i = w_0 + \sum_{j=1}^{d} w_j x_{ij} - y_i$. The second objective function is the sum of square Euclidean distances to the hyperplane; i.e., $\sum_{i=1}^{n} r_i^2$, where r_i is the Euclidean distance between point (x_i, y_i) to the hyperplane $f(x) = w_0 + \sum_{j=1}^{d} w_j x_j$.

a) (5 points) Derive a gradient descent algorithm to find the parameters of the model that minimizes the sum of squared errors.

b) (20 points) Derive a gradient descent algorithm to find the parameters of the model that minimizes the sum of squared distances.

c) (20 points) Implement both algorithms and test them on 5 datasets. Datasets can be randomly generated, as in class, or obtained from resources such as UCI Machine Learning Repository. Compare the solutions to the closed-form (maximum likelihood) solution derived in class and find the R^2 in all cases on the same dataset used to fit the parameters; i.e., do not implement cross-validation.