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.

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.

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?

Carry out all steps manually and show all your calculations.

Solution

a) The prior probabilities are as follows: $p(+) = 1/2, p(-) = 1/2$
Likelyhood of the feature x_1: $p(x_1 = 1|+) = p(x_1 = -1|+) = p(x_1 = 1|-) = p(x_1 = -1|-) = 1/2$
Likelyhood of the feature x_2: $p(x_2 = 1|+) = p(x_2 = -1|+) = p(x_2 = 1|-) = p(x_2 = -1|-) = 1/2$
Posterior probabilities:

$$p(+|X_1) = \frac{p(X_1|+)p(+)}{p(X_1|+)p(+) + p(X_1|-)p(-)} = \frac{p(x_1 = -1, x_2 = -1|+)p(+)}{p(x_1 = -1, x_2 = -1|+)p(+) + p(x_1 = -1, x_2 = -1|-)p(-)}$$

Now, using the Naive Bayes' assumption.

$$p(+|X_1) = \frac{p(x_1 = -1|+)p(x_2 = -1|+)p(+)}{p(x_1 = -1|+)p(x_2 = -1|+)p(+) + p(x_1 = -1|-)p(x_2 = -1|-)p(-)} = 1/2$$

Similarly,

$$p(-|X_1) = \frac{p(x_1 = -1|-)p(x_2 = -1|-)p(-)}{p(x_1 = -1|+)p(x_2 = -1|+)p(+) + p(x_1 = -1|-)p(x_2 = -1|-)p(-)} = 1/2$$

$$p(+|X_2) = \frac{p(x_1 = -1|+)p(x_2 = 1|+)p(+)}{p(x_1 = -1|+)p(x_2 = 1|+)p(+) + p(x_1 = -1|-)p(x_2 = 1|-)p(-)} = 1/2$$

$$p(-|X_2) = \frac{p(x_1 = -1|-)p(x_2 = 1|-)p(-)}{p(x_1 = -1|+)p(x_2 = 1|+)p(+) + p(x_1 = -1|-)p(x_2 = 1|-)p(-)} = 1/2$$

$$p(+|X_3) = \frac{p(x_1 = 1|+)p(x_2 = -1|+)p(+)}{p(x_1 = 1|+)p(x_2 = -1|+)p(+) + p(x_1 = 1|-)p(x_2 = -1|-)p(-)} = 1/2$$
$$p(-|X_3) = \frac{p(x_1 = 1|-)p(x_2 = -1|-)p(-)}{p(x_1 = 1|+)p(x_2 = -1|+)p(+) + p(x_1 = 1|-)p(x_2 = -1|-)p(-)} = 1/2$$
$$p(+|X_4) = \frac{p(x_1 = 1|+)p(x_2 = 1|+)p(+)}{p(x_1 = 1|+)p(x_2 = 1|+)p(+) + p(x_1 = 1|-)p(x_2 = 1|-)p(-)} = 1/2$$
$$p(-|X_4) = \frac{p(x_1 = 1|-)p(x_2 = 1|-)p(-)}{p(x_1 = 1|+)p(x_2 = 1|+)p(+) + p(x_1 = 1|-)p(x_2 = 1|-)p(-)} = 1/2$$

Since, this classifier classifies both the correct outcomes and the incorrect outcomes with the probability of 0.5, therefore the accuracy of the classifier can be said to be as 0.5 (or 50%).

b) The posterior probabilities and the likelyhoods of the old features would remain the same from the previous part. Likelyhoods of the new features:

$$p(+1|+) = 1/2, p(+1|-) = 1/2,$$

$$p(x_1x_2 = 1|+) = 0, p(x_1x_2 = -1|+) = 1, p(x_1x_2 = 1|-) = 1, p(x_1x_2 = -1|-) = 0,$$
$$p(x_1^2 = 1|+) = 1, p(x_1^2 = -1|+) = 0, p(x_1^2 = 1|-) = 1, p(x_1^2 = -1|-) = 0,$$
$$p(x_2^2 = 1|+) = 1, p(x_2^2 = -1|+) = 0, p(x_2^2 = 1|-) = 1, p(x_2^2 = -1|-) = 0.$$

Using the Naive Bayes' Assumption:

$$p(X_1|+) = p(+1|+)p(x_1 = -1|+)p(x_2 = -1|+)p(x_1x_2 = 1|+)p(x_1^2 = 1|+)p(x_2^2 = 1|+) = 0$$

Therefore, $p(+|X_1) = 0$ and $p(-|X_1) = 1$ Similarly, we get

$$p(X_2|+) = p(+1|+)p(x_1 = -1|+)p(x_2 = 1|+)p(x_1x_2 = -1|+)p(x_1^2 = 1|+)p(x_2^2 = 1|+) = 1/32$$

Therefore, $p(+|X_2) = 1$ and $p(-|X_2) = 0$. Similarly,

$$p(X_3|+) = p(+1|+)p(x_1 = 1|+)p(x_2 = -1|+)p(x_1x_2 = -1|+)p(x_1^2 = 1|+)p(x_2^2 = 1|+) = 1/32$$

Therefore, $p(+|X_3) = 1$ and $p(-|X_3) = 0$. Similarly,

$$p(X_4|+) = p(+1|+)p(x_1 = 1|+)p(x_2 = 1|+)p(x_1x_2 = 1|+)p(x_1^2 = 1|+)p(x_2^2 = 1|+) = 0$$

Therefore, $p(+|X_4) = 0$ and $p(-|X_4) = 1$.
This classifier makes the correct guess in each case therefore its accuracy is 1 (or 100%).

c) The posterior probabilities and the likelyhoods of the old features would remain the same from the previous part. Likelyhoods of the new features:

$$p(-x_1x_2 = 1|+) = 1, p(-x_1x_2 = -1|+) = 0, p(-x_1x_2 = 1|-) = 0, p(-x_1x_2 = -1|-) = 1$$

Its likelyhood would always have the same value as the likelyhood of the feature x_1x_2, and since that likelyhood is always either 1 or 0, adding this extra feature would not change any of the likelyhoods or the posterior probabilities as shown below:

$$p(X_1|+) = p(+1|+)p(x_1 = -1|+)p(x_2 = -1|+)p(x_1x_2 = 1|+)p(x_1^2 = 1|+)p(x_2^2 = 1|+)p(-x_1x_2 = -1|+) = 0$$
$$p(X_2|+) = p(+1|+)p(x_1 = -1|+)p(x_2 = 1|+)p(x_1x_2 = -1|+)p(x_1^2 = 1|+)p(x_2^2 = 1|+)p(-x_1x_2 = 1|+) = 1/32$$
$$p(X_3|+) = p(+1|+)p(x_1 = 1|+)p(x_2 = -1|+)p(x_1x_2 = -1|+)p(x_1^2 = 1|+)p(x_2^2 = 1|+)p(-x_1x_2 = 1|+) = 1/32$$
$$p(X_4|+) = p(+1|+)p(x_1 = 1|+)p(x_2 = 1|+)p(x_1x_2 = 1|+)p(x_1^2 = 1|+)p(x_2^2 = 1|+)p(-x_1x_2 = -1|+) = 0$$

Hence, $p(+|X_1) = 0, p(-|X_1) = 1, p(+|X_2) = 1, p(-|X_2) = 0, p(+|X_3) = 1, p(-|X_3) = 0, p(+|X_4) = 0, p(-|X_4) = 1$. And as we can see all these probabilities remain the same as before.
Therefore the accuracy of this classifier is 1 (or 100%) as well.

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.

Solution

a) We know that the boundary of the decision surface can be calculated by $P(Y = 1|\mathbf{x}) = P(Y = 0|\mathbf{x})$. Therefore, we get

$$\frac{p(\mathbf{x}|Y = 1)P(Y = 1)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|Y = 1)P(Y = 0)}{p(\mathbf{x})}$$

Therefore,

$$p(\mathbf{x}|Y = 1) = p(\mathbf{x}|Y = 0)$$

$$\frac{1}{(2\pi)^{1/2}|\mathbf{\Sigma}_1|^{1/2}} \cdot e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_1)^T\mathbf{\Sigma}_1^{-1}(\mathbf{x}-\mathbf{m}_1)} = \frac{1}{(2\pi)^{1/2}|\mathbf{\Sigma}_2|^{1/2}} \cdot e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_2)^T\mathbf{\Sigma}_2^{-1}(\mathbf{x}-\mathbf{m}_2)}$$

Canceling the constants on both sides since they're the same and then taking log of both sides (we can do this because both sides are positive), we get

$$(\mathbf{x} - \mathbf{m}_1)^T\mathbf{\Sigma}_1^{-1}(\mathbf{x} - \mathbf{m}_1) = (\mathbf{x} - \mathbf{m}_2)^T\mathbf{\Sigma}_2^{-1}(\mathbf{x} - \mathbf{m}_2)$$

Taking $\mathbf{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}$ and taking the values of all the other matrices we get,

$$(x_1 - 1)^2 + (x_2 - 2)^2 = (x_1 - 6)^2 + (x_2 - 3)^2$$

$$5x_1 + x_2 = 20$$

b) We know that the boundary of the decision surface can be calculated by $P(Y = 1|\mathbf{x}) = P(Y = 0|\mathbf{x})$. Therefore, we get

$$\frac{p(\mathbf{x}|Y = 1)P(Y = 1)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|Y = 1)P(Y = 0)}{p(\mathbf{x})}$$

Taking $P(Y = 1) = p$ and $P(Y = 0) = 1 - p$. Therefore,

$$p\frac{1}{(2\pi)^{1/2}|\mathbf{\Sigma}_1|^{1/2}} \cdot e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_1)^T\mathbf{\Sigma}_1^{-1}(\mathbf{x}-\mathbf{m}_1)} = (1-p)\frac{1}{(2\pi)^{1/2}|\mathbf{\Sigma}_2|^{1/2}} \cdot e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_2)^T\mathbf{\Sigma}_2^{-1}(\mathbf{x}-\mathbf{m}_2)}$$

Canceling common constants and taking log of both sides, we get

$$(-1/2)(\mathbf{x} - \mathbf{m}_1)^T\mathbf{\Sigma}_1^{-1}(\mathbf{x} - \mathbf{m}_1) + \log p = (-1/2)(\mathbf{x} - \mathbf{m}_2)^T\mathbf{\Sigma}_2^{-1}(\mathbf{x} - \mathbf{m}_2) + \log 1 - p$$

Taking $\mathbf{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}$ and taking the values of all the other matrices we get,

$$\frac{(x_1 - m_{01})^2 + (x_2 - m_{02})^2}{2\sigma^2} + \log p = \frac{(x_1 - m_{11})^2 + (x_2 - m_{12})^2}{2\sigma^2} + \log 1 - p$$

$$x_1^2 - 2x_1 m_{01} + m_{01}^2 + x_2^2 - 2x_2 m_{02} + m_{02}^2 + 2\sigma^2 \log p = x_1^2 - 2x_1 m_{11} + m_{11}^2 + x_2^2 - 2x_2 m_{12} + m_{12}^2 + 2\sigma^2 \log 1 - p$$

Therefore we get the equation of the decision surface as

$$2x_1(m_{11} - m_{01}) + 2x_2(m_{12} - m_{02}) + m_{01}^2 + m_{02}^2 - m_{11}^2 - m_{12}^2 + 2\sigma^2 \log \frac{p}{1-p} = 0$$

c) We know that the boundary of the decision surface can be calculated by $P(Y = 1|\mathbf{x}) = P(Y = 0|\mathbf{x})$. Therefore, we get

$$\frac{p(\mathbf{x}|Y = 1)P(Y = 1)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|Y = 1)P(Y = 0)}{p(\mathbf{x})}$$

Taking $P(Y = 1) = p$ and $P(Y = 0) = 1 - p$. Therefore,

$$p\frac{1}{(2\pi)^{1/2}|\mathbf{\Sigma}_1|^{1/2}} \cdot e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_1)^T \mathbf{\Sigma}_1^{-1}(\mathbf{x}-\mathbf{m}_1)} = (1-p)\frac{1}{(2\pi)^{1/2}|\mathbf{\Sigma}_2|^{1/2}} \cdot e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_2)^T \mathbf{\Sigma}_2^{-1}(\mathbf{x}-\mathbf{m}_2)}$$

Canceling common constants and taking log of both sides, we get the equation of the decision boundary in vector form as follows

$$(-1/2)(\mathbf{x}-\mathbf{m}_1)^T \mathbf{\Sigma}_1^{-1}(\mathbf{x}-\mathbf{m}_1) + \log p - (1/2)\log|\mathbf{\Sigma}_1| = (-1/2)(\mathbf{x}-\mathbf{m}_2)^T \mathbf{\Sigma}_2^{-1}(\mathbf{x}-\mathbf{m}_2) + \log(1-p) - (1/2)\log|\mathbf{\Sigma}_2|$$

In both of the previous parts the decision surface was just a plane of d-dimensions (2 in those cases) because its equation was just the linear combination of all the features (i.e. x_1 and x_2). In this part though, the equation of the decision surface is not just a linear combination but rather terms of degree 2 may exist as well (depending on the exact values of the covariance matrices). Hence the surface will be a quadratic plane.

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.

Solution

a) Considering the $x_{i0} = 1 \forall i \in [1, n]$. The objective function:

$$C = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(\sum_{j=0}^{d} w_j x_{ij} - y_i \right)^2$$

The partial derivative of the objective function with respect to some $w_k \forall k \in [0, d]$:

$$\frac{\partial C}{\partial w_k} = 2 \sum_{i=1}^{n} \left(\sum_{j=0}^{d} w_j x_{ij} - y_i \right) x_{ik}$$

Let $W = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix}$ and $\Delta = \begin{bmatrix} \frac{\partial C}{\partial w_0} \\ \frac{\partial C}{\partial w_1} \\ \vdots \\ \frac{\partial C}{\partial w_1} \end{bmatrix}$.

Therefore the gradient descent update becomes: $W^{(t+1)} = W^{(t)} - \eta \Delta^{(t)}$.Now, we just have to initialize the value of $W^{(0)}$ and choose an appropriate η, then we can just run the update until convergence, and this is the gradient descent algorithm.

b) Considering the $x_{i0} = 1 \forall i \in [1, n]$. The objective function:

$$C = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} \left(\frac{\sum_{j=0}^{d} w_j x_{ij} - y_i}{\sqrt{\sum_{j=0}^{d} w_J^2 + y_i^2}} \right)^2$$

The partial derivative of the objective function with respect to some $w_k \forall k \in [0, d]$:

$$\frac{\partial C}{\partial w_k} = \sum_{i=1}^{n} \frac{2(\sum_{j=0}^{d} w_j x_{ij} - y_i) x_{ik} (\sum_{j=0}^{d} w_j^2 + y_i^2) - 2 w_k (\sum_{j=0}^{d} w_j x_{ij} - y_i)^2}{(\sum_{j=0}^{d} w_j^2 + y_i^2)^2}$$

$$\frac{\partial C}{\partial w_k} = 2 \sum_{i=1}^{n} \frac{(\sum_{j=0}^{d} w_j x_{ij} - y_i) \left(x_{ik} (\sum_{j=0}^{d} w_j^2 + y_i^2) - w_k (\sum_{j=0}^{d} w_j x_{ij} - y_i) \right)}{(\sum_{j=0}^{d} w_j^2 + y_i^2)^2}$$

Let $W = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix}$ and $\Delta = \begin{bmatrix} \frac{\partial C}{\partial w_0} \\ \frac{\partial C}{\partial w_1} \\ \vdots \\ \frac{\partial C}{\partial w_1} \end{bmatrix}$.

Therefore the gradient descent update becomes: $W^{(t+1)} = W^{(t)} - \eta \Delta^{(t)}$.Now, we just have to initialize the value of $W^{(0)}$ and choose an appropriate η, then we can just run the update until convergence, and this is the gradient descent algorithm.