CSE 575: Statistical Machine Learning Assignment #1

Instructor: Prof. Hanghang Tong
Out: Jan. 22nd, 2016; Due: Feb. 12th, 2016

Submit electronically, using the submission link on Blackboard for Assignment #1, a file named `yourFirstName-yourLastName.pdf` *containing your solution to this assignment (a .doc or .docx file is also acceptable, but .pdf is preferred).*

1 Bayes Classifier [20 points]

Prove that the Bayes classifier is the optimal, i.e., the expected risk of a Bayes classifier is minimal among all the possible classifiers. You only need to show this for binary classifiers.

Solution: For any given example x, the risk of the Bayes classifier is $r_{bayes} = min(q_1(x), q_2(x))$, which is smaller than or equal to the risk of any other classifier, where $q_i(x)$ $(i = 1, 2) = p(y = i|x)$.

Therefore, the expected risk of the Bayes classifier must be the smallest among all possible (binary) classifiers.

2 Parameter Estimation [20 points]

For this question, assume that $x_1, \dots, x_N \in \mathbb{R}$ are i.i.d samples drawn from the same underlying distribution. Assume that the underlying distribution is Gaussian $N(\mu, \sigma^2)$.

1. (5 points) What is the MLE estimator of μ?

   Solution. $\hat{\mu}_{MLE} = \frac{\sum_{i=1}^{N} x_i}{N}$

2. (5 points) Is your MLE estimator of μ a random variable? **Explain.**

   Solution. Yes. $\hat{\mu}_{MLE}$ is a function of $x_1, \dots, x_N$. Each of them is a random variable. So $\hat{\mu}_{MLE}$ is also a random variable.

3. (5 points) Let $\hat{\mu}_{MLE}$ denote the MLE estimator of μ. Please prove that $\hat{\mu}_{MLE}$ is unbiased. **Hint: The bias of an estimator of the parameter μ is defined to be the difference between the expected value of the estimator and μ.**

   Solution. $E(\hat{\mu}_{MLE}) = E(\frac{\sum_{i=1}^{N} x_i}{N}) = \frac{1}{N}\sum_{i=1}^{N} E(x_i) = \mu$. So $\hat{\mu}_{MLE}$ is unbiased.

4. (5 points) If the true value of μ is known, then the MLE estimator of σ^2 is as follows.

   $$\hat{\sigma}^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

   Please prove that $\hat{\sigma}^2_{MLE}$ is unbiased. Notice that this estimator is different from the one we introduced in class due to the fact that we already know the true value of μ.

   Solution. $E(\hat{\sigma}^2_{MLE}) = E(\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2) = \frac{1}{N}\sum_{i=1}^{N} E(x_i - \mu)^2 = \mu$. So $\hat{\sigma}^2_{MLE}$ is unbiased.

3 Naive Bayes Classifier [20 points]

Given the training data set in Figure 1, we want to train a binary classifier, with (1) the last column being the class label (i.e., whether or not to enjoy the sport); and (2) each column of X being a binary feature.

X						Y
Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

Figure 1: Training Data Set for Naive Bayes Classifiers

1. (5 points) How many independent parameters are there in your Naive Bayes classifier? What are they? Justifiy your answer.

   Solutions: (1) $P(y=1)$ (i.e., enjoy sports), (2) $P(x_1 = sunny|y = i)$ $(i = 1, 0)$, (3) $P(x_2 = warm|y = i)$ $(i = 1, 0)$, (4) $P(x_3 = normal|y = i)$ $(i = 1, 0)$, (5) $P(x_4 = strong|y = i)$ $(i = 1, 0)$, (6) $P(x_5 = warm|y = i)$ $(i = 1, 0)$, (7) $P(x_7 = same|y = i)$ $(i = 1, 0)$. 13 independent parameters in total.

2. (10 points) What are your estimations for these parameters? (say using standard MLE).

   Solutions: (1) $P(y = 1) = 3/4$ (i.e., enjoy sports), (2) $P(x_1 = sunny|y = 1) = 1$ and $P(x_1 = sunny|y = 0) = 0$, (3) $P(x_2 = warm|y = 1) = 1$ and $P(x_2 = warm|y = 0) = 0$, (4) $P(x_3 = normal|y = 1) = 1/3$ and $P(x_3 = normal|y = 0) = 0$, (5) $P(x_4 = strong|y = 1) = 1$ and $P(x_4 = strong|y = 0) = 1$, (6) $P(x_5 = warm|y = 1) = 2/3$ and $P(x_5 = warm|y = 0) = 1$, (7) $P(x_7 = same|y = 1) = 2/3$ and $P(x_7 = same|y = 0) = 0$.

3. (5 points) Now, given a new (test) example $x = (sunny, warm, high, strong, cool, change)$, what is $P(y = 1|x)$? Which class label will the naive Bayes classifer assign to this example? Justify your answer.

   Solutions: $P(x|y = 1)P(y = 1) = P(x_1 = sunny|y = 1)P(x_2 = warm|y = 1)P(x_3 = high|y = 1)P(x_4 = strong|y = 1)P(x_5 = cool|y = 1)P(x_6 = change|y = 1)P(y = 1) = 1 \times 1 \times 2/3 \times 1 \times 1 \times 1/3 \times 1/3 \times 3/4 = 1/18$. $P(x|y = 0)P(y = 0) = 0$. Therefore, $P(y = 1|x) = \frac{P(x|y=1)P(y=1)}{P(x|y=1)P(y=1)+P(x|y=0)P(y=0)} = 1$. The assigned label will be $y = 1$.

4 1NN-Classifier Decision Boundary [20 points]

Given two training data points as shown in Figure **??**, what is the decision boundary of 1NN classifier if we use L_2 distance (10 points)? What will be the decision boundary if we use L_∞ distance instead (10 points)?Justify your answer.

Solutions: (1) the y-axis. (2) shown in the following figure.

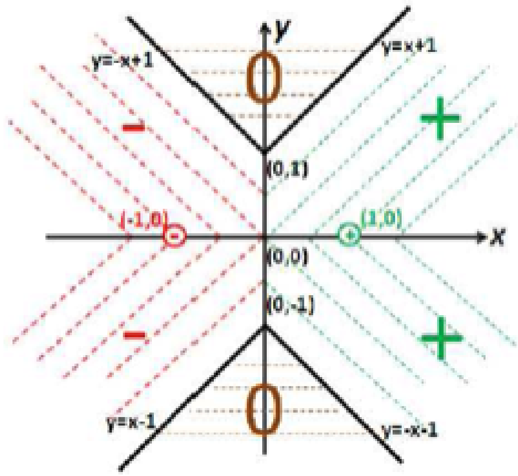

Figure 2: Decision Boundary of 1NN Classifiers with L_∞ Distance.

5 The decision boundary for 1NN (i.e., *1*-Nearest Neighbors Classifier) [20 points]

For each of the following figures, we are given a few data points in the 2-d space, each of which is labeled as either '+' or '-'. Draw the decision boundary for 1NN, assuming we use L_2 distance.

Solution. 5 pts for each case, and no partial credits for each case.

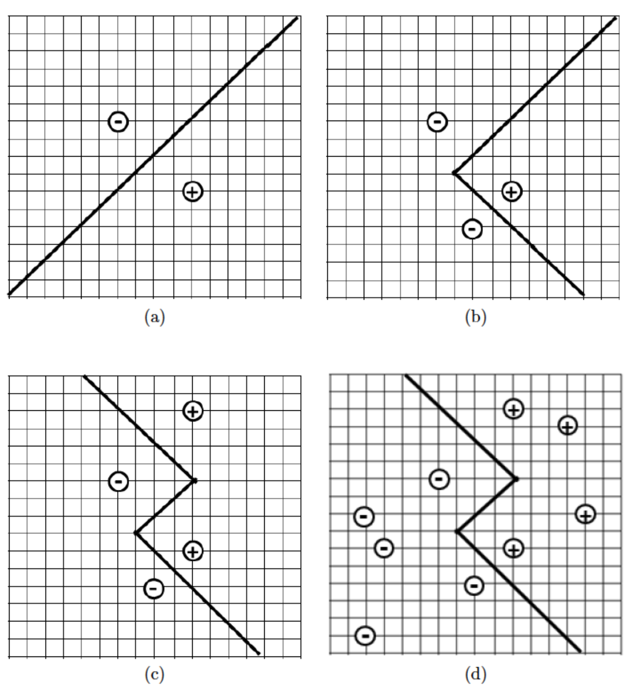

Figure 3: Training Data Set for 1NN Classifiers