PRML Homework Assignment II

(Due on Saturday, May 30, 2020.)

Please submit to our TA–Siwen Liu before 8:30 am.

For every 5 minutes beyond the deadline, 5 points will be deducted from your score.

Part I

Problem 1: Probability Theory (20 points)

Suppose we have two bags each containing black and white balls. Bag A contains 50 black balls, and twice as many white balls as black balls. Bag B contains 30 white balls, and 3 times as many black balls as white balls.

(a) (5 points) Suppose we give Bag A to a robot who is programmed to randomly select one ball from the bag, and then put the ball back to the bag after recording its color. After the robot performs the task 3 times, what is the probability that it records 2 black balls and 1 white ball?

(b) (5 points) Suppose we replace Bag A with Bag B, and give the bag to the same robot. After the robot performs the task 3 times, what is the probability that it records 2 black balls and 1 white ball?

(c) (10 points) Suppose we randomly select one bag, and give it to this robot. After the robot performs the task 3 times, it records 2 black balls and 1 white ball. What is the probability that these balls are from Bag A?

Problem 2: Bayesian Decision Theory (20 points)

Consider a two-category one-dimensional problem. Suppose the class-conditional densities and the priors of the two classes are known (i.e. $p(x|\omega_1)$, $p(x|\omega_2)$, $P(\omega_1)$ and $P(\omega_2)$).

(a) (5 points) Suppose the decision rule is: Decide ω_1 if $x > \theta$; otherwise decide ω_2. Suppose θ is not the optimal decision boundary, as shown in Figure 1. Explain that the average probability of error for this rule is given by

$$P(error) = P(\omega_1) \int_{-\infty}^{\theta} p(x|\omega_1)dx + P(\omega_2) \int_{\theta}^{\infty} p(x|\omega_2)dx.$$

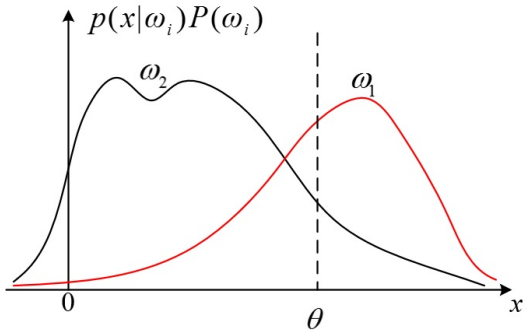

Figure 1: An example of two-category one-dimensional classification.

(b) (5 points) If we change the decision rule to "maximizing a posterior (MAP)", then show that the average probability of error is given by

$$P(error) = 1 - \int P(\omega_{max}|x)p(x)dx,$$

where $P(\omega_{max}|x) \geq P(\omega_i|x)$ for all $i, i = 1, 2$.

(c) (2 points) Describe a situation when the two decision rules are equivalent.

(d) (8 points) Suppose we extend this problem to a m-category one-dimensional case. Show that $P(\omega_{max}|x) \geq 1/m$ and $P(error) \leq (m-1)/m$.

Problem 3: Maximum-likelihood Estimation and Bayesian Parameter Estimation (35 points)

Maximum-likelihood estimation (MLE) and Bayesian parameter estimation (BPE) can be applied to estimate the prior probability as well. Let samples be drawn by successive, independent selections of a state of nature ω_i with unknown probability $P(\omega_i)$. Let $z_{ik} = 1$ if the state of nature for the kth sample is ω_i and $z_{ik} = 0$ otherwise.

(a) (2 points) We first reformulate this problem with the notations used in the lectures. For example, the sample set $\mathcal{D} = \{x_1, \ldots, x_n\} = \{z_{i1}, \ldots, z_{in}\}$. Please represent θ and $p(\mathcal{D}|\theta)$ with $\{z_{i1}, \ldots, z_{in}\}$ and $P(\omega_i)$.

(b) (6 points) For MLE approach, show that

$$p(\mathcal{D}|\theta) = \prod_{k=1}^{n} \theta^{x_k}(1-\theta)^{1-x_k}.$$

(Hint: $p(x|\theta)$ in (d)).

(c) (10 points) For MLE approach, show that the best estimate for θ is

$$\theta = \frac{1}{n}\sum_{k=1}^{n} x_k.$$

(d) (5 points) For BPE approach, we assume that

$$p(x|\theta) = \begin{cases} \theta & x = 1 \\ 1-\theta & x = 0 \end{cases} \Leftrightarrow p(x|\theta) = \theta^x(1-\theta)^{(1-x)}(x = 0, 1),$$

and $p(\theta)$ follows uniform distribution (i.e. $p(\theta) \sim\ U(0, 1)$). Please derive $p(\theta|\mathcal{D})$. (Hint: $\int_0^1 \theta^m(1-\theta)^n d\theta = \frac{m!n!}{(m+n+1)!}$)

(e) (5 points) For BPE approach, show that

$$p(x|\mathcal{D}) = \frac{1}{n+2} \cdot \frac{(x + \sum_{k=1}^{n} x_k)!(n + 1 - x - \sum_{k=1}^{n} x_k)!}{(\sum_{k=1}^{n} x_k)!(n - \sum_{k=1}^{n} x_k)!}$$

(f) (5 points) What is the effective Bayesian estimate for θ? (Hint: What does $p(x = 1)$ stand for?)

(g) (2 points) Compare the estimation results of MLE and BPE approaches and describe your findings.

Part II

Problem 1: Maximizing-A-Posterior (MAP) Decision Rule and Maximum-likelihood Estimation (MLE) (35 points)

In Chapter 3, we learnt how to design an optimal classifier based on MAP decision rule if we knew the prior probabilities and the class-conditional densities of all categories (denoted as $\omega_i, i = 1, 2, ..., c.$). However, in practical pattern recognition applications we rarely have this kind of complete knowledge. Therefore, we introduced parameter estimation approaches which illustrated how to use data samples to estimate the unknown probabilities and probability densities.

Let us consider a three-category classification problem, and assume that the three categories have equal priors. The sampled data are stored in three files, data1.txt, data2.txt, and data3.txt (Note: the files can be downloaded from the shared folder of the QQ group, hw2_partII.zip). The first file contains the samples of class 1, the second file contains the samples of class 2 and so forth. Each sample has two dimensions and each file contains

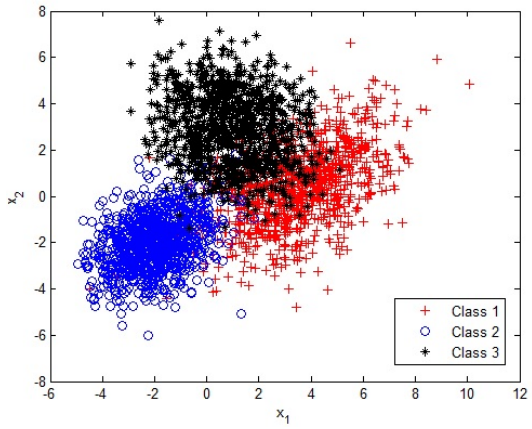

Figure 2: The first 1000 samples of each category.

2000 samples. In the following exercises, we will use part of the samples to train three MAP classifiers, one for each category, and then evaluate the classifiers with the rest of the samples.

(a) (0 points) Warming up. Plot the first 1000 samples of each category. Your result should be similar to Figure 2.

(b) (10 points)Since $P(\omega_1) = P(\omega_2) = P(\omega_3)$, the posterior probabilities $P(\omega_i|\mathbf{x}), i = 1, 2, 3$, are determined by the class-conditional densities $p(\mathbf{x}|\omega_i), i = 1, 2, 3$. Hence, we first estimate/train $p(\mathbf{x}|\omega_i), i = 1, 2, 3$ with MLE approach. Assume that

$$p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i).$$

Use the first 1000 samples of each category to estimate $p(\mathbf{x}|\omega_i)$ and plot the resulting class-conditional densities.

(c) (10 points) Use MAP decision rule to classify the rest of the samples and show the rate of misclassification in each category.

(d) (10 points) If we use the first 500 samples of each category for training, and the rest for testing. Repeat (b) and (c), and show the rate of misclassification in each category.

(e) (5 points) Describe and interpret your findings by comparing the results of (c) and (d).

Problem 2: Parzen Window Estimation and k-Nearest Neighbor (k-NN) Estimation (40 points)

Let us revisit the three-category classification in Problem 1, but utilize non-parametric methods to perform classification this time. We assume that the three categories have equal priors. The sampled data are stored in three files, data1.txt, data2.txt, and data3.txt (Note: the files can be downloaded from the shared folder of the QQ group, hw2_partII.zip). The first file contains the samples of class 1, the second file contains the samples of class 2 and so forth. Each sample has two dimensions and each file contains 2000 samples.

(a) (10 points) In Chapter 5, one important conclusion about parzen window estimation method is "$p_n(\mathbf{x})$ is actually the average of n normalized window functions, each centered at $\mathbf{x}_i$" where $p_n(\mathbf{x})$ denotes the estimate of the true density function and $\mathbf{x}_i$ denotes the i-th sample ($1 \leq i \leq n$). Alternatively, the parzen window estimate $p_n(\mathbf{x})$ can be expressed as:

$$p_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\delta_n(\mathbf{x}-\mathbf{x}_i),$$

where $\delta_n(\cdot)$ denotes the normalized window function. Suppose the window function is a multivariate Gaussion function,

$$\delta_n(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}}\exp[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})],$$

where $\boldsymbol{\mu} = \mathbf{0}$ and $\mathbf{\Sigma} = \mathbf{I}$ ($\mathbf{I}$ is the identity matrix). Use the first 1000 samples of each category to obtain $p_n(\mathbf{x})$ for each category and plot the parzen window estimate of the density function of each category.

(b) (10 points) Use the results of (a) and Maximizing-A-Posterior (MAP) decision rule to classify the rest of the samples and show the misclassification rate of each category.

(c) (10 points) In Chapter 5, we also introduced how to use k-NN method to estimate the posterior distribution. Suppose we are given the first 1000 samples of each category as training samples, and asked to classify the rest of the samples using NN classifier (i.e. k=1). Show the misclassification rate of each category.

(d) (5 points) Repeat (c) when $k = 10$. (Hint: $P(\omega_i|\mathbf{x}) = k_{\omega_i}/k$, where k_{ω_i} is the number of samples labeled ω_i within the k nearest neighbors of $\mathbf{x}$)

(e) (5 points) Describe and interpret your findings by comparing the results of (c) and (d).

Submission Guidelines for Part II

1. Create a zip file, YourID_HW1.zip, which includes your source code and a short report that clearly illustrates the required results. The report has to be in the format of pdf or doc. In addition, please verify that all the files can be successfully extracted from the zip file before submission.