Homework #3 – Parametric Methods

CAP 5638, Pattern Recognition (Fall 2015), Department of Computer Science, Florida State University

Points: 50 Due: Wednesday, October 7, 2015

Submission: Hardcopy (including programs) is required and is due at the beginning of the class on the due date.

Problem 1 (10 points) Problem 1 (**parts (a) and (b) only**), Chapter 3 of the textbook

1. Let x have an exponential density

$$p(x|\theta) = \begin{cases} \theta e^{-\theta x} & x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Plot $p(x|\theta)$ versus x for $\theta = 1$. Plot $p(x|\theta)$ versus θ, $(0 \leq \theta \leq 5)$, for $x = 2$.

(b) Suppose that n samples $x_1, ..., x_n$ are drawn independently according to $p(x|\theta)$. Show that the maximum likelihood estimate for θ is given by

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

Problem 2 (10 points) Problem 3, Chapter 3 of the textbook

3. Maximum likelihood methods apply to estimates of prior probabilities 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) Show that

$$P(z_{i1}, \ldots, z_{in}|P(\omega_i)) = \prod_{k=1}^{n} P(\omega_i)^{z_{ik}}(1 - P(\omega_i))^{1-z_{ik}}.$$

(b) Show that the maximum likelihood estimate for $P(\omega_i)$ is

$$\hat{P}(\omega_i) = \frac{1}{n}\sum_{k=1}^{n} z_{ik}.$$

Interpret your result in words.

Problem 3 (15 points) Problem 7, Chapter 3 of the textbook

7. Show that if our model is poor, the maximum likelihood classifier we derive is not the best — even among our (poor) model set — by exploring the following example. Suppose we have two equally probable categories (i.e., $P(\omega_1) = P(\omega_2) = 0.5$). Further, we know that $p(x|\omega_1) \sim N(0, 1)$ but *assume* that $p(x|\omega_2) \sim N(\mu, 1)$. (That is, the parameter θ we seek by maximum likelihood techniques is the mean of the second distribution.) Imagine however that the *true* underlying distribution is $p(x|\omega_2) \sim N(1, 10^6)$.

(a) What is the value of our maximum likelihood estimate $\hat{\mu}$ in our poor model, given a large amount of data?

(b) What is the decision boundary arising from this maximum likelihood estimate in the poor model?

(c) Ignore for the moment the maximum likelihood approach, and use the methods from Chap. ?? to derive the Bayes optimal decision boundary given the *true* underlying distributions — $p(x|\omega_1) \sim N(0,1)$ and $p(x|\omega_2) \sim N(1,10^6)$. Be careful to include all portions of the decision boundary.

(d) Now consider again classifiers based on the (poor) model assumption of $p(x|\omega_2) \sim N(\mu, 1)$. Using your result immediately above, find a *new* value of μ that will give lower error than the maximum likelihood classifier.

(e) Discuss these results, with particular attention to the role of knowledge of the underlying model.

Problem 4 (5 points) Problem 10, Chapter 3 of the textbook (Hint: think about the bias and variance.)

10. Suppose we employ a novel method for estimating the mean of a data set $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$: we assign the mean to be the value of the first point in the set, i.e., $\mathbf{x}_1$.

(a) Show that this method is unbiased.

(b) State why this method is nevertheless highly undesirable.

Problem 5 (10 points) Suppose that the prior distribution of θ and the parametric form (a uniform distribution) remain the same as in the example given in Section 3.5 in the textbook, compute first the Bayesian estimation of θ and then the estimated class conditional p(x | D) for D={3, 9, 7}. You need to specify the Bayesian estimation and the class conditional fully (i.e., you need to specify the functions with all required constants). Then plot the class conditional from 0 to 10.

Extra Credit Problem

Problem 6 (7 points) Problem 11, Chapter 3 of the textbook; you only need to show the univariate case.

11. One measure of the difference between two distributions in the same space is the *Kullback-Leibler divergence* of Kullback-Leibler "distance":

$$D_{KL}(p_1(\mathbf{x}), p_2(\mathbf{x})) = \int p_1(\mathbf{x}) \ln \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} d\mathbf{x}.$$

(This "distance," does not obey the requisite symmetry and triangle inequalities for a metric.) Suppose we seek to approximate an arbitrary distribution $p_2(\mathbf{x})$ by a normal $p_1(\mathbf{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Show that the values that lead to the smallest Kullback-Leibler divergence are the obvious ones:

$$\begin{aligned} \boldsymbol{\mu} &= \mathcal{E}_2[\mathbf{x}] \\ \boldsymbol{\Sigma} &= \mathcal{E}_2[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t], \end{aligned}$$

where the expectation taken is over the density $p_2(\mathbf{x})$.