E1 213 Pattern Recognition and Neural Networks

Practice Problems: Set 3

1. Suppose a class conditional density is normal with mean μ and variance 10. Suppose we assume that the class conditional density is normal with unknown mean and variance 1. Suppose we have large amount of data and do a ML estimate for μ. Would the estimated mean be close to the true mean? What can be said about the error rate of the resulting Bayes classifier?

Answer: Yes, the estimated μ would be close to its true value because the sample size is large. However, the final class conditional density we use is wrong. The error rate of resulting Bayes classifuer can be very bad (depending on what the other class conditional density is).

2. The Bernoulli random variable that we considered takes values 0 and 1. Sometimes we want binary random variables that take values $+1$ or -1. Suppose X is a random variable that takes values $+1$ and -1 with probabilities p and $(1-p)$ respectively. Write the mass function of X with p as a parameter and derive the ML estimate for p. Can you justify the final answer in terms of 'sample mean'? Denote by μ the mean of X. Write the mass function of X with μ as the parameter.

Answer: The mass function can be written as

$$f(x|p) = p^{\frac{1+x}{2}} (1-p)^{\frac{1-x}{2}},\ x \in \{-1,\ +1\}$$

Note that $EX = p + (-1)(1-p) = 2p - 1$.
By differentiating likelihood function, one can show

$$\hat{p}_{ML} = \frac{n + \sum_{i=1}^{n} x_i}{2n}$$

This can be seen to be reasonable in terms of sample mean estimator for expectation:

$$\hat{p}_{ML} = \frac{n + \sum_{i=1}^{n} x_i}{2n} = \frac{1}{2}\left(1 + \frac{1}{n}\sum_{i=1}^{n} x_i\right)$$

Mass function with μ as parameter is

$$f(x|\mu) = \left(\frac{1+\mu}{2}\right)^{\frac{1+x}{2}} \left(\frac{1-\mu}{2}\right)^{\frac{1-x}{2}},\ x \in \{-1,\ +1\}$$

3. Suppose X is a discrete random variable with mass function

$$f(x \mid p) = (1-p)^{x-1}p,\ x = 1, 2, \cdots$$

where p (with $0 < p < 1$) is the parameter. (This is geometric random variable). What is the ML estimate for p? Suppose we want a Bayesian estimate for p. What is the conjugate prior? What would be the MAP estimate? Is $\sum_i x_i$ a sufficient statistic for p?

Answer: The likelihood function is

$$L(p|\mathcal{D}) = \prod_{i=1}^{n}(1-p)^{x_i-1}p = p^n(1-p)^{-n+\sum x_i}$$

The log-likelihood is

$$l(p|\mathcal{D}) = n\ln(p) - n\ln(1-p) + \sum_i x_i \ln(1-p)$$

By differentiating this and equating to zero, we get

$$\hat{p}_{ML} = \frac{n}{\sum_{i=1}^{n} x_i}$$

The likelihood function has the form $p^c(1-p)^d$. Conjugate prior is Beta distribution. If prior is Beta(a, b), then we get

$$f(p|\mathcal{D}) \propto p^n(1-p)^{-n+\sum x_i}\ p^{a-1}(1-p)^{b-1}$$

Now, using the formula for mode of beta density, we get

$$\hat{p}_{MAP} = \frac{n+a-1}{\sum_{i=1}^{n} x_i + a + b - 2}$$

If $a = b = 1$ then $\hat{p}_{ML} = \hat{p}_{MAP}$.

It is easy to guess from the ML estimate that $\sum_i x_i$ is a sufficient statistic for p. You can easily verify it.

4. We want to estimate θ, which is the probability of heads of a coin. The data consists of N tosses of which N_1 are heads. Suppose we want a Bayesian estimate. Suppose our prior density is

$$\begin{aligned} f(\theta) &= 0.5 \text{ if } \theta = 0.5 \\ &= 0.5 \text{ if } \theta = 0.6 \\ &= 0 \text{ otherwise} \end{aligned}$$

Find the MAP estimate of θ.

Answer: I hope it is intuitively clear that

$$\hat{\theta}_{MAP} = \begin{cases} 0.5 & \text{if } \frac{N_1}{N} \leq 0.55 \\ 0.6 & \text{otherwise} \end{cases}$$

Let $x_1, \cdots, x_N$ be the data where x_i is 1 if that toss is head and zero otherwise. So, we know $\sum x_i = N_1$. Now the posterior is given by

$$\begin{aligned} f(\theta|\mathcal{D}) &\propto \prod_{i=1}^{N} f(x_i|\theta)\ f(\theta) \\ &= \prod_{i=1}^{N} (\theta)^{x_i}(1-\theta)^{1-x_i}\ f(\theta) \\ &= \theta^{N_1}(1-\theta)^{N-N_1}\ f(\theta) \end{aligned}$$

With the given prior, we know the posterior is zero unless $\theta = 0.5$ or 0.6. Hence, the MAP estimate would be 0.5 if

$$(0.5)^{N_1}(0.5)^{N-N_1} \ \geq\ (0.6)^{N_1}(0.4)^{N-N_1}$$

(otherwise it would be 0.6). Simplifying this, you get the MAP estimate given earlier. (Please do this simplification).

5. Consider a 2-class problem where class conditional densities are normal. We have a large number of feature vectors in our training set. Hoewever, we do not have class labels for any of these feature vectors. Can we still learn the class conditional densities and implement a Bayes classifier?

Answer: Yes, we can estimate a mixture density. See your first assignment.

6. Let Z be a K-dimensional random vector with joint mass function given by

$$f(\mathbf{z}) = \prod_{i=1}^{K} \rho_i^{z_i}$$

where $\mathbf{z} = [z_1\ \cdots,\ z_K]^T$ with $z_i \in \{0, 1\}$ and $\sum_i z_i = 1$. The ρ_i are the parameters. (Note that $\rho_i \geq 0$ and $\sum_i \rho_i = 1$). Let X be a random variable whose conditional density conditioned on $\mathbf{z}$ is given by

$$f(x \mid \mathbf{z}) = \prod_{i=1}^{K} \left(\phi(x \mid \mu_i, \sigma_i)\right)^{z_i}$$

where $\phi(x \mid \mu_i, \sigma_i)$ is a guassian density function with mean μ_i and variance σ_i^2. Show that the (marginal) density of X is a mixture of K Gaussians.

Answer: Note that $f(\mathbf{z})$ is non-zero only for K number of K-dimensional vectors, namely, $[1\ 0 \cdots 0]^T, [0,\ 1\ 0 \cdots 0]^T, \cdots, [0\ 0 \cdots 0\ 1]^T$.

Now we have (note that x is continuous and $\mathbf{z}$ is discrete)

$$\begin{aligned} f(x) &= \sum_{\mathbf{z}} f(x \mid \mathbf{z}) f(\mathbf{z}) \\ &= \sum_{\mathbf{z}} \prod_{i=1}^{K} (\rho_i\ \phi(x \mid \mu_i, \sigma_i))^{z_i} \\ &= \rho_1\ \phi(x \mid \mu_1, \sigma_1) + \cdots + \rho_K\ \phi(x \mid \mu_K, \sigma_K) \end{aligned}$$

7. We considered the EM algorithm in the context of ML estimation in the sense that it is meant for maximizing the log likelihood. Suppose we want to obtain the MAP estimate and want to make use of hidden or latent variables. As in the notation used in class, let $\mathbf{x}$ be the observed data and let $\mathbf{z}$ be the hidden data. We want to maximize $\ln(f(\theta \mid \mathbf{x}))$ which is the log of the posterior density. Show that we can use the EM algorithm where the E-step computes the same $Q(\theta, \theta^{(k)})$ as earlier but now the M-step maximizes, over θ, $Q(\theta, \theta^{(k)}) + \ln(f(\theta))$ where $f(\theta)$ is the prior density.

Answer: We have

$$\ln(f(\theta \mid \mathbf{x})) \propto \ln(f(\mathbf{x}|\theta)) + \ln(f(\theta))$$

So, to maximize the posterior, we can maximize the RHS above and this cane rewritten (as in the analysis of EM algorithm) as

$$\ln(f(\mathbf{x}, \mathbf{z}|\theta)) - \ln(f(\mathbf{z}|\mathbf{x}, \theta)) + \ln(f(\theta))$$

Now we take expectation of this with respect to conditional density of $\mathbf{z}$ conditioned on $\mathbf{x}$ and $\theta^{(k)}$. This gives us (using the same symbols as in the EM algorithm),

$$Q(\theta, \theta^{(k)}) + \ln(f(\theta)) - R(\theta, \theta^{(k)})$$

Now it is clear that if we maximize $Q(\theta, \theta^{(k)}) + \ln(f(\theta))$ over θ then we would be increasing $\ln(f(\theta \mid \mathbf{x}))$.