CS181 Practice Questions: Generative Classification

1. **MLE for Probabilistic Classification (Bishop 4.9)**

Consider a generative classification model for K classes defined by prior class probabilities $p(\mathcal{C}_k) = \pi_k$ and general class-conditional densities $p(\phi|\mathcal{C}_k)$ where ϕ is the input feature vector. Suppose we are given a training data set $\{\phi_n, \mathbf{t}_n\}$ where $n = 1, \ldots, N$ and $\mathbf{t}_n$ is a binary target vector of length K that uses the 1-of-K coding scheme, so that it has components $t_{nj} = I_{jk}$ if pattern n is from class $\mathcal{C}_k$. Assuming that the data points are drawn independently from this model, show that the maximum-likelihood solution for the prior probabilities is given by:

$$\pi_k = \frac{N_k}{N},$$

where N_k is the number of data points assigned to class $\mathcal{C}_k$.

2. **MLE for Gaussian Probabilistic Classification (Bishop 4.10)**

Consider the classification model of the previous exercise and now suppose that class-conditional densities are given by Gaussian distributions with a shared covariance matrix, so that

$$p(\phi|\mathcal{C}_k) = \mathcal{N}(\phi|\mu_k, \mathbf{\Sigma})$$

Show that the maximum likelihood solution for the mean of the Gaussian distribution for class $\mathcal{C}_k$ is given by

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} t_{nk} \phi_n$$

which represents the mean of those feature vectors assigned to class $\mathcal{C}_k$. Similarly, show that the maximum likelihood solution for the shared covariance matrix is given by

$$\mathbf{\Sigma} = \sum_{k=1}^{K} \frac{N_k}{N} \mathbf{S}_k$$

where

$$\mathbf{S}_k = \frac{1}{N_k} \sum_{n=1}^{N} t_{nk} (\phi_n - \mu_k)(\phi_n - \mu_k)^\mathsf{T}.$$

Thus $\mathbf{\Sigma}$ is given by a weighted average of the covariances of the data associated with each class, in which the weighting coefficients are given by the prior probabilities of the classes.

3. **Gaussian Decision Boundaries (Murphy 4.21)**

Consider a two-class, generative classification model where $p(x \mid \mathcal{C}_j) = \mathcal{N}(x \mid \mu_j, \sigma_j^2)$. Let $\mu_1 = 0$, $\sigma_1^2 = 1$, $\mu_2 = 1$, $\sigma_2^2 = 10^6$, and let the class priors be $p(\mathcal{C}_1) = p(\mathcal{C}_2) = \frac{1}{2}$. Find the decision region $\mathcal{R}_1 = \{x \mid p(\mathcal{C}_1 \mid x) \geq p(\mathcal{C}_2 \mid x)\}$.
Hint: find the solutions of the equation $p(x|\mu_1, \sigma_1^2) = p(x|\mu_2, \sigma_2^2)$, and recall that to solve a quadratic equation $ax^2 + bx + c = 0$, we use

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

4. **Logistic Regression vs LDA/QDA (Murphy 4.20)**

Suppose we train the following binary classifiers via maximum likelihood:

- GaussI: A generative classifier, where the class conditional densities are Gaussian, with both covariance matrices set to $\boldsymbol{I}$, i.e., $p(\boldsymbol{x}|\mathcal{C}_j) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{I})$.
- GaussX: same as GaussI, but the covariance matrices are unconstrained, i.e., $p(\boldsymbol{x}|\mathcal{C}_j) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$.
- LinLog: A logistic regression model with linear features.
- QuadLog: A logistic regression model, using linear and quadratic features.

Now suppose that after training, we evaluate the conditional log-likelihood of the *training set* under each model. That is, for each model we evaluate

$$L(\theta, M) = \sum_{n=1}^{N} \ln p(t_n|\phi(\boldsymbol{x}_n), \theta),$$

where θ are our MLE parameters and M the model in question. For each of the following pairs of models, indicate which model will have lower (or equal) $L(\theta, M)$, or indicate that no such statement can be made. Explain your answers.

a. GaussI, LinLog

b. GaussX, QuadLog

c. LinLog, QuadLog

d. GaussI, QuadLog

Finally, is it in general true that a classifier that gives a higher $L(\theta, M)$ on the training set will have fewer classification errors on the training set?