CS 479/679 Pattern Recognition
Sample Midterm Exam

1. **[20 pts]** True/False Questions – To get credit, you **must** give brief reasons for your answers.

   T F The Bayes rule is always an optimum classification rule in the sense that it minimizes the average probability error.

   True ONLY under the assumption that $p(x/\omega i)$ and $P(\omega i)$ have been modelled/estimated correctly; in this case, the Bayes rule is optimum in the sense that it minimizes the average probability error.

   T F Adding more features will in general improve classification accuracy.

   False since adding more features will in general deteriorate classification accuracy due to the "curse" of dimensionality (i.e., the number of training data required increases exponentially with the number of features).

   T F A causal relationship typically exists between correlated events.

   False since correlated events are not necessarily causally related (i.e., correlations between two events can be caused by a third factor, called confounder, which affects both events).

   T F A linear discriminant function could optimally separate the data between two classes when the features are uncorrelated.

   False since the covariance matrix (diagonal in this case) for each class might be different (i.e., features can have different variances for each class). This is Case 3 (i.e., non-linear discriminant).

   T F Given a set of N random variables $X=\{X_1, X_2, \ldots, X_N\}$, the pdf of any single random variable X_i can be computed from the joint pdf of X.

   True by applying marginalization with respect to all other variables except X_i

2. **[15 pts]]** Short answer questions.

a. Why do simpler models typically perform better than complex models in pattern recognition?

Complex models "memorize" the training data, therefore, they generalize poorly.

b. What are the fundamental similarities/differences between Maximum A-Posteriori parameter estimation (MAP) and Bayesian parameter estimation (BE)?

Both MAP and BE assume known $p(\theta)$

MAP: fixed point estimate

BE: estimates a distribution

c. Under what conditions would the optimal decision boundary between two classes, each modeled by a Gaussian distribution, **not** pass from the midpoint of the line joining the means of the distributions?

When $P(\omega_1) \neq P(\omega_2)$ the decision boundary does <u>not</u> pass through the midpoint for cases I, II (case III could be equal or not equal)

d. How could a matrix A be diagonalized? Are all matrices A diagonalizable? Explain.

$P^{-1} A P = \Lambda$

P: columns are the eigenvectors of A

Λ: diagonal, eigenvalues of A

- Not all matrices are diagonalizable
- Eigenvectors of A must be linearly independent

3. **[15 pts]** Consider a c-class classification problem; under what conditions would the optimal classifier be equivalent to (a) the minimum distance classifier? (b) the Mahalanobis distance classifier?

In both cases, we assume that each class is modelled by a multivariate Gaussian distribution.

(a) $\Sigma_i = \sigma^2 I$

assuming equal priors

(b) $\Sigma_i = \Sigma$

assuming equal priors

4. **[15 pts]** Consider the following probability distribution:

$$p(x) = \begin{cases} (\theta + 1)x^{\theta} & for\ 0 \le x \le 1 \\ 0 & otherwise \end{cases}$$

Given n points $x_1, x_2, ..., x_n$ sampled from the above distribution, derive a formula for the maximum likelihood estimate of ϑ.

$$p(D/\theta) = \prod_{i=1}^{n} p(x_i/\theta)$$

$$\ln p(D/\theta) = \sum_{i=1}^{n} \ln p(x_i/\theta) = \sum_{i=1}^{n} \ln\left[(\theta+1)x_i^{\theta}\right] =$$

$$= \sum_{i=1}^{n} \ln(\theta+1) + \sum_{i=1}^{n} \ln(x_i^{\theta}) =$$

$$= n\ln(\theta+1) + \theta \sum_{i=1}^{n} \ln(x_i)$$

Take derivative and set equal to zero:

$$\frac{\partial \ln p(D/\theta)}{\partial \theta} = \frac{n}{\theta+1} + \sum_{i=1}^{n} \ln(x_i) = 0$$

$$\Rightarrow \frac{n}{\theta+1} = -\sum_{i=1}^{n} \ln(x_i) \Rightarrow \theta+1 = -\frac{n}{\sum_{i=1}^{n} \ln(x_i)}$$

$$\Rightarrow \boxed{\theta = -1 - \frac{n}{\sum_{i=1}^{n} \ln(x_i)}}$$

5. **[20 pts]** Discuss the main ideas of **the parameter estimation** techniques presented in class and compare them with each other. Are their solutions always different? Explain.

Maximum Likelihood Estimation

It assumes that the parameters are fixed but unknown. It maximizes the likelihood **p(D/θ)** where D is the training data and θ the parameters to be estimated.

Maximum A-Posteriori Estimation:

It assumes that the parameters are fixed and have a known distribution p(θ). It maximizes the likelihood **p(D/θ)p(θ)** where D is the training data and θ the parameters to be estimated.

Bayesian Estimation:

It assumes that the parameters are random variables with a known distribution p(θ). Using the data D, it converts p(θ) to p(θ/D). It estimates a distribution:

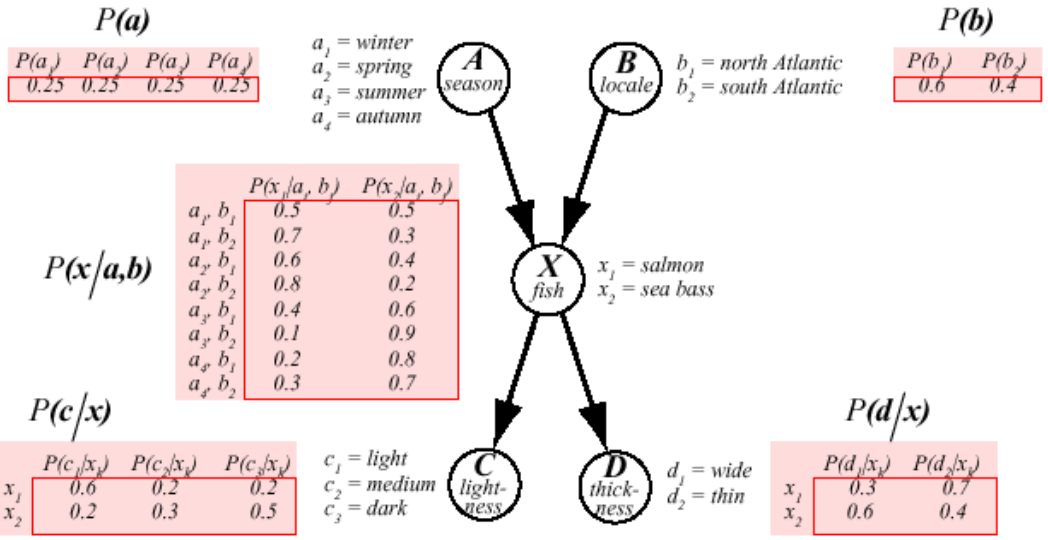
$$p(\mathbf{x} / D) = \int p(\mathbf{x} / \boldsymbol{\theta})\, p(\boldsymbol{\theta} / D) d\boldsymbol{\theta}$$

Comparison

MLE/MAP are simpler than BE and their solution is simpler to interpret. They also have lower computational requirements compared to BE. The BE solution might not be of the form (model) assumed.

The BE solution is different from the MLE solution when the dataset is not very large. In theory, the BE solution converges to the MLE solution as n→∞

6. **[15 pts]** Using the Bayesian network below, answer the following question: suppose all we know is that a fish is thin, has medium lightness, and was caught in south Atlantic; what season is it now, most likely?



We need to compute $P(a_i/d_2, c_2, b_2)$

$$P(a_i/d_2, c_2, b_2) = \frac{P(a_i, d_2, c_2, b_2)}{P(d_2, c_2, b_2)} \sim \sum_{a_j} P(a_i, b_2, x_j, c_2, d_2$$

$$P(a_1/d_2, c_2, b_2) = \sum_j P(a_1, b_2, x_j, c_2, d_2) =$$

$$= P(a_1, b_2, x_1, c_2, d_2) + P(a_1, b_2, x_2, c_2, d_2) =$$

$$P(a_1)P(b_2)P(x_1/a_1, b_2)P(c_2/x_1)P(d_2/x_1)$$

$$P(a_1)P(b_2)P(x_2/a_1, b_2)P(c_2/x_2)P(d_2/x_2)$$

$$= 0.0134$$

Similarly $P(a_2/d_2, c_2, b_2) = 0.0136$ ← highest most likely spring!

$P(a_3/d_2, c_2, b_2) = 0.0122$

$P(a_4/d_2, c_2, b_2) = 0.0126$