Homework 5

1. ***Allowing Classifiers to Reject Data***

Suppose as usual $X_1, \ldots, X_n$ are hypothesis events that partition a sample space Ω, and $Y_1, \ldots, Y_m$ are observation events that also partition Ω. So far we have designed classifiers that measure which Y_i occurs and estimate which X_j occurred. As usual, we have $x \in \mathbb{R}^n$ is the prior pmf and $A \in \mathbb{R}^{m \times n}$ is the matrix of transition probabilities,

$$A_{ij} = \text{Prob}(Y_i \,|\, X_j)$$

and from these we can find

- the *a-posteriori probabilities* $B_{ij} = \text{Prob}(X_j \,|\, Y_i)$
- the *marginal pmf* $y_i = \text{Prob}(Y_i)$
- and the *joint distribution* $J_{ij} = \text{Prob}(Y_i \cap X_j)$

We would like to add another option to the classifier, so that it can choose to *reject* the data; i.e., make no decision. The key idea that if the cost for rejection is low compared with the cost for misclassifying, then rejecting may be a better option.

To formulate this, we need to generalize our notion of classification to include more general *decision problems*. We think about the estimator as measuring which Y_i occurs, and then deciding to take one of p decisions $\{D_1, \ldots, D_p\}$. We will use $p = n + 1$, so that when $j \leq n$, taking decision D_j corresponds to estimating X_j, and decision D_{n+1} means reject the data.

We will now use an estimator $f_{\text{est}} : \{1, \ldots, m\} \to \{1, \ldots, p\}$, and alternatively specify it by a matrix $K \in \mathbb{R}^{m \times p}$ where

$$K_{ij} = \begin{cases} 1 & \text{if } j = f_{\text{est}}(i) \\ 0 & \text{otherwise} \end{cases}$$

We will also define the error matrix $E \in \mathbb{R}^{p \times n}$ by

$$E_{jk} = \text{probability that decision } D_j \text{ is taken and event } X_k \text{ occurs}$$

(a) Show that $E = K^T J$.

(b) Now we will assign costs via the matrix $C \in \mathbb{R}^{p \times n}$ where

$$C_{jk} = \text{cost of taking decision } D_j \text{ when } X_k \text{ occurs}$$

Show that the expected cost is

$$\mathrm{E}\, C = \text{trace}(K^T J C^T)$$

(c) Let $W = JC^T$. Show that K is optimal if $K\mathbf{1} = \mathbf{1}$ and

$$K_{ij} = 1 \qquad \text{implies} \qquad W_{ij} \leq W_{ik} \text{ for all } k$$

that is, we pick the entry in row i of K corresponding to the smallest entry of row i of W.

(d) Now we can use this to formulate the classifier with rejection. Define the cost

$$C_{jk} = \begin{cases} 0 & \text{if } j = k \\ \lambda_r & \text{if } j = n + 1 \\ \lambda_m & \text{otherwise} \end{cases}$$

Here we interpret $\lambda_r \geq 0$ as the cost for choosing to reject the data, and $\lambda_m \geq 0$ as the cost for misclassifying which of hypothesis events X_i occurred. Show that K is an optimal classifier if $K\mathbf{1} = \mathbf{1}$ and for all $j = 1, \ldots, n$ we have

$$K_{ij} = 1 \qquad \text{implies} \qquad B_{ij} \geq B_{ik} \text{ for all } k, \text{ and } B_{ij} \geq 1 - \frac{\lambda_r}{\lambda_m}$$

That is, to estimate X_j when Y_i occurs, not only must the a-posteriori probability B_{ij} be greater than all other a-posteriori probabilities B_{ik}, but also it must be greater than the threshold $1 - \frac{\lambda_r}{\lambda_m}$.

(e) Suppose we have a radar with conditional pmfs

$$p^{\text{nothing}} = \frac{1}{100}\begin{bmatrix}4 & 7 & 11 & 12 & 17 & 16 & 13 & 9 & 6 & 4 & 1 & 0 & 0 & 0 & 0\end{bmatrix}$$
$$p^{\text{something}} = \frac{1}{100}\begin{bmatrix}0 & 0 & 0 & 0 & 2 & 6 & 8 & 10 & 15 & 15 & 15 & 13 & 9 & 5 & 2\end{bmatrix}$$

Here X_1 is the event that nothing is visible, and X_2 is the event that something is visible, and Y_i is the event that the radar returns i pings (at least one ping is always returned.) Then

$$\text{Prob}(Y_i \,|\, X_1) = p_i^{\text{nothing}} \qquad \text{Prob}(Y_i \,|\, X_2) = p_i^{\text{something}}$$

with prior probabilities

$$\text{Prob}(X_1) = 0.6 \qquad \text{Prob}(X_2) = 0.4;$$

Find the classifier, without rejection, that minimizes the probability of error.

(f) Now assume $\lambda_r = 0.2$ and $\lambda_m = 1$. Find the classifier with rejection that minimizes the Bayes risk defined by the cost function in part (d). What are the corresponding optimal cost and probability of a correct estimate?

(g) What is the probability of a *wrong estimate*, excluding those cases where the classifier rejects the data?

(h) How does using $\lambda_r > 0$ affect the probability of a correct estimate? And the probability of a wrong estimate (excluding rejections) ?

(i) Explain what happens when $\lambda_r = 0$

(j) Explain what happens when $\lambda_r \geq \lambda_m$

2. ***Gaussians and Confidence Regions***

In this question we will investigate the properties of confidence ellipsoids, and the relationship between the confidence ellipsoid and the marginal confidence interval.

Suppose $x \sim \mathcal{N}(0, \Sigma)$ where

$$\Sigma = \begin{bmatrix}2 & 0.8\\ 0.8 & 1\end{bmatrix}$$

(a) For each p in $\{0.1, 0.2, \ldots, 0.9\}$ plot the ellipsoid C_p corresponding to confidence level p. That is, with probability p the vector x will land in C_p.

(b) Collect $10,000$ samples of x, and count the fraction of them that lie within C_p, for each p.

(c) What is the marginal distribution of x_1?

(d) Find the 90% confidence interval $[-c_1, c_1]$ for x_1. Make a plot which shows it (as vertical bars) on top of the ellipsoid $C_{0.9}$.

(e) Using the same data as before, count what fraction of points has $x_1 \in [-c_1, c_1]$.

(f) The *projection* of $C_{0.9}$ onto the x_1 axis is an interval $[-c_2, c_2]$. Find c_2 and plot it also.

(g) Now a more general question about ellipsoid projections. Suppose

$$E = \left\{ x \in \mathbb{R}^n \mid x^T \Sigma^{-1} x \leq 1 \right\}$$

where $\Sigma \in \mathbb{R}^{n \times n}$. Partition $x = \begin{bmatrix}x_1\\ x_2\end{bmatrix}$ where $x_1 \in \mathbb{R}^r$. The projection of E onto the plane

$$P = \left\{ x \in \mathbb{R}^n \mid x_2 = 0 \right\}$$

is an ellipsoid. Find this ellipsoid, i.e, find Q such that it is

$$\left\{ x_1 \in \mathbb{R}^r \mid x_1^T Q x_1 \leq 1 \right\}$$

Hint: completion of squares will help here.

(h) Show that the ratio c_2/c_1 is a constant; that is, that it does not depend on the covariance Σ. What is this ratio? Repeat your computation in part (d) for the case when $\Sigma = I$ to verify this.

3. ***The χ^2_n distribution***

Suppose $x : \Omega \to \mathbb{R}$ and $x \sim \mathcal{N}(0, \sigma^2)$. Let $y = x^2/\sigma^2$. Show that y has pdf given by the χ^2_1 distribution.