Homework 5 Solutions

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 $\lambda_r \geq \lambda_m$

Solution.

(a) This part follows the proof in the notes. Notice that D_j is an event, so we can write

$$\begin{aligned}
E_{jk} &= \text{Prob}(D_j \cap X_k) \\
&= \sum_{i=1}^{m} \text{Prob}(D_j \cap X_k \cap Y_i) \qquad \text{since the } Y_i \text{ partition } \Omega \\
&= \sum_{i=1}^{m} \text{Prob}\Big(f_{\text{est}}(i) = j \text{ and } Y_i \text{ and } X_k\Big) \\
&= \sum_{i=1}^{m} \text{Prob}\Big(\bigcup\{\, Y_p \,|\, f_{\text{est}}(p) = j \,\} \cap Y_i \text{ and } X_k\Big)
\end{aligned}$$

Now we use the fact that

$$\bigcup\{\, Y_p \,|\, f_{\text{est}}(p) = j \,\} \cap Y_i = \begin{cases} Y_i & \text{if } K_{ij} = 1 \\ \emptyset & \text{otherwise} \end{cases}$$

Therefore we have

$$\begin{aligned}
E_{jk} &= \sum_{i=1}^{m} K_{ij} \,\text{Prob}(Y_i \cap X_k) \\
&= \sum_{i=1}^{m} K_{ij} J_{ik}
\end{aligned}$$

That is, $E = K^T J$ as required.

(b) The expected cost is

$$\begin{aligned}\mathrm{E}\, C &= \sum_{j=1}^{p}\sum_{k=1}^{n} C_{jk}\,\mathrm{Prob}(D_j \cap X_k) \\ &= \sum_{j=1}^{p}\sum_{k=1}^{n} C_{jk}E_{jk} \\ &= \mathrm{trace}(EC^T) \\ &= \mathrm{trace}(K^T J C^T)\end{aligned}$$

(c) We need to find a K with binary entries, with exactly one nonzero entry in each row. Further we would like to minimize

$$\mathrm{trace}(K^T W) = \sum_{i=1}^{m}\sum_{j=1}^{p} K_{ij}W_{ij}$$

Hence for each i, we look for the smallest entry of row i of W, and set the corresponding entry of K to be one. More precisely, we want

$$K_{ij} = 1 \quad \text{only if} \quad W_{ij} \le W_{ik} \quad \text{for all } k$$

In the case where two entries in row i of W are the same, then we can choose either and achieve the same cost.

(d) We would like to find the minimum element of row i of W where

$$W = JC^T$$

Since

$$J_{ij} = y_i B_{ij}$$

we can equivalently find the minimum element in row i of Q_{ij}, where $Q = BC^T$. That is, row i of W is just row i of Q scaled by y_i. Also notice that

$$C^T = \begin{bmatrix}\lambda_m(\mathbf{1}\mathbf{1}^T - I) & \lambda_r\mathbf{1}\end{bmatrix}$$

and since $B\mathbf{1} = \mathbf{1}$ we have

$$Q = \begin{bmatrix}\lambda_m(\mathbf{1}\mathbf{1}^T - B) & \lambda_r\mathbf{1}\end{bmatrix}$$

The entries of row i of Q are therefore

$$\lambda_m(1 - B_{i1}), \lambda_m(1 - B_{i2}), \ldots, \lambda_m(1 - B_{in}), \lambda_r$$

Hence element Q_{ij} is the smallest in the row only if

$$B_{ij} \ge B_{ik} \;\text{ for all } k, \text{ and } B_{ij} \ge 1 - \frac{\lambda_r}{\lambda_m}$$

(e) The optimal classifier without rejection is a threshold classifier as follows.

$$f_{\mathrm{est}}(i) = \begin{cases} 1 & \text{if } 1 \le i \le 8 \\ 2 & \text{otherwise}\end{cases}$$

(f) The optimal classifier is

$$f_{\mathrm{est}}(i) = \begin{cases} 1 & \text{for } 1 \le i \le 6 \\ 3 & \text{for } 7 \le i \le 10 \\ 2 & \text{for } 11 \le i \le 15\end{cases}$$

Optimal cost is $\mathbf{trace}(K^T J C^T) = 0.1148$. The error matrix is

$$E = \begin{bmatrix}0.402 & 0.032 \\ 0.006 & 0.176 \\ 0.192 & 0.192\end{bmatrix}$$

The probability of a correct estimate is $E_{11} + E_{22} = 0.578$.

(g) The probability of a wrong estimate, excluding those cases where the classifier rejects the data, is $E_{21} + E_{12} = 0.038$.

(h) As λ_r increases, the penalty for making a rejction increases compared to the penalty for misclassification. Hence the classifier will make decisions which are not guaranteed to be correct. Hence both the probability of a correct estimate as well as the probability of an incorrect classification increase.

(i) If $\lambda_r = 0$ then there is no cost for rejection, and the classifier will only make a decision when it is guaranteed to be correct. Such a classifier is optimal.

However, there is another optimal classifier, and this is one that always rejects. This is because the cost of rejection and the cost of a correct classification are the same.

(j) When $\lambda_r \geq \lambda_m$ there will never be a rejection. Hence the classifier in this case is equivalent to the one in part (e).

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.

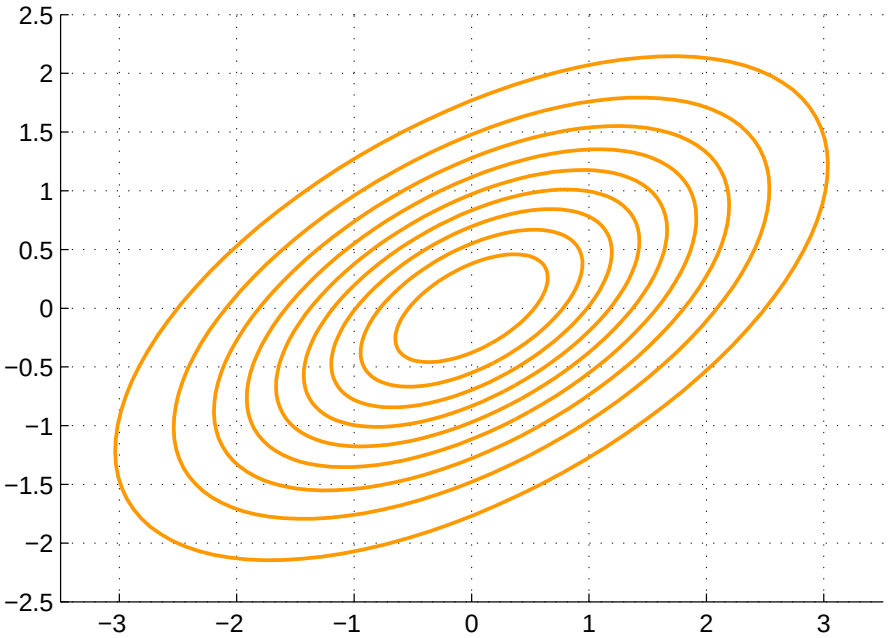
(h) Show that the ratio c_2/c_1 is a constant; that is, that it does not depend on the covariance Σ. What is this ratio? Repeat your computation in part (d) for the case when $\Sigma = I$ to verify this.

Solution.

(a) Since x has zero mean, $x^T\Sigma^{-1}x$ has a χ^2_2 distribution. The confidence ellipsoid corresponding to a probability p is given by

$$C_p = \left\{x \in \mathbb{R}^2 | x^T\Sigma^{-1}x \leq F^{-1}_{\chi^2_2}(p)\right\}$$

where $F_{\chi^2_2}$ is the cumulative distribution function for a χ^2_2 distribution. The ellipsoids are plotted below.



(b) We can generate $x \sim \mathcal{N}(0, \Sigma)$ in Matlab via

```
x=sqrtm(cov)*randn(2,1
```

where `cov` is the matrix Σ. Then x lies in the confidence ellipsoid C_p if and only if

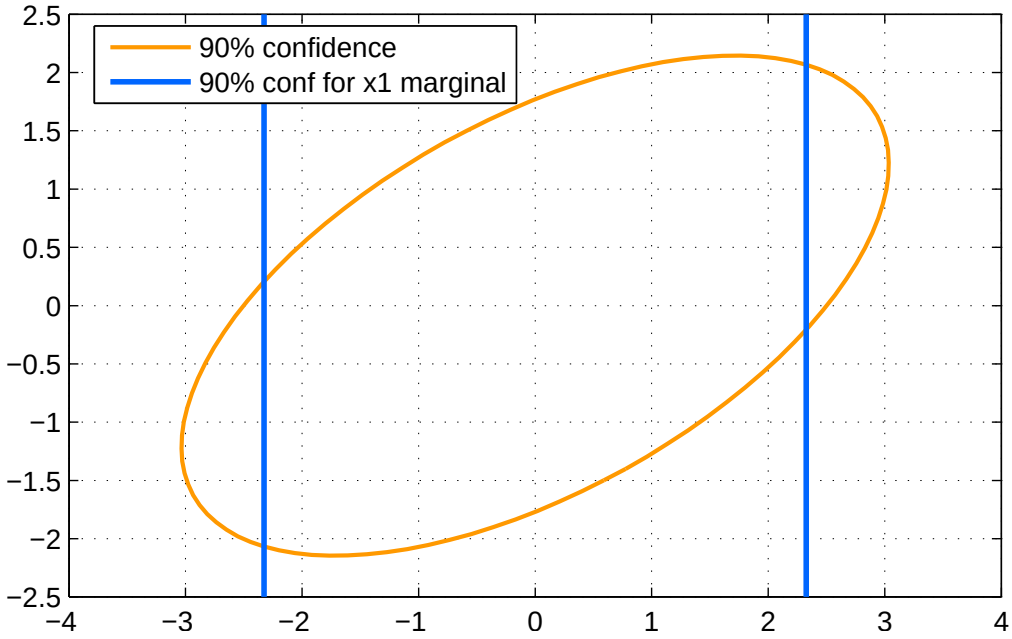
$$x^T\Sigma^{-1}x \leq F^{-1}_{\chi^2_2}(p)$$

(c) The marginal distribution of x_1 is $\mathcal{N}(0, 2)$.

(d) We can use the χ^2_1 distribution to find c_1.

$$c_1 = \sqrt{\Sigma_{11} * F^{-1}_{\chi^2_1}(0.9)}$$

which gives $c_1 \approx 2.33$. The confidence interval and ellipsoid is below.

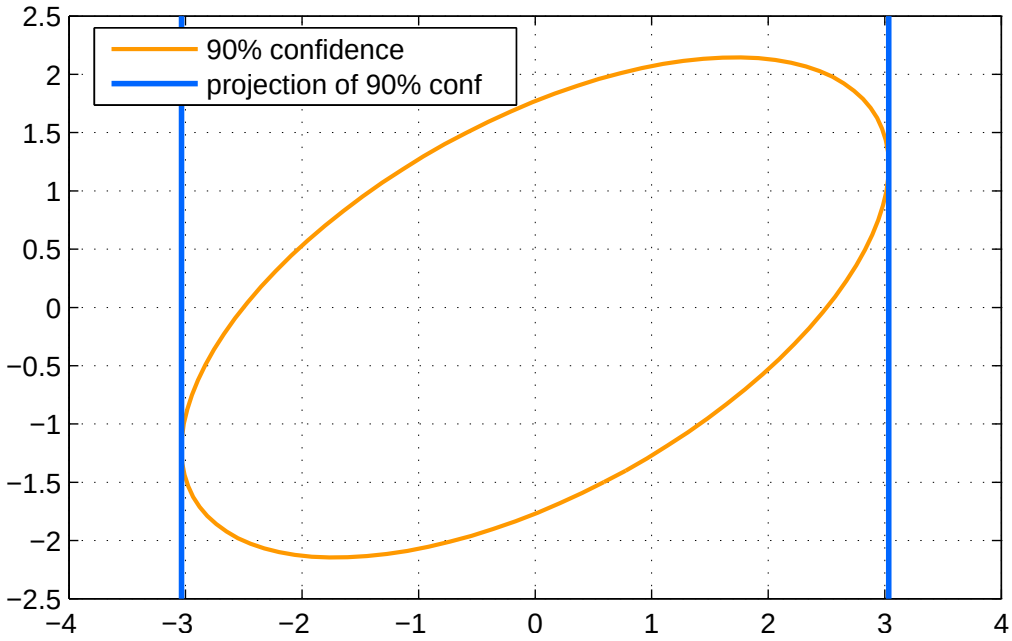


(e) Counting gives approximately 90% of points.

(f) We find (using the formula below) the half-width is

$$c_2 = \sqrt{\Sigma_{11} * F_{\chi_2^2}^{-1}(0.9)} \approx 3.03$$

and this is plotted below.



(g) The required projection is the set

$$E_{\text{proj}} = \left\{ x_1 \in \mathbb{R}^r \,|\, \text{there exist } x_2 \in \mathbb{R}^{n-r} \text{ such that } \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \leq 1 \right\}$$

This we can write as

$$E_{\text{proj}} = \left\{ x_1 \in \mathbb{R}^r \mid \min_{x_2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \leq 1 \right\}$$

Now we can use completion of squares to give

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_1^T \Sigma_{11}^{-1} x_1 + (x_2 - \Sigma_{21}\Sigma_{11}^{-1}x_1)^T S^{-1}(x_2 - \Sigma_{21}\Sigma_{11}^{-1}x_1)$$

where $S = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{21}$. Also since $\Sigma \succeq 0$, we have $S \succeq 0$, and hence

$$\min_{x_2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_1^T \Sigma_{11}^{-1} x_1$$

Therefore the projection of E is given by

$$E_{\text{proj}} = \left\{ x_1 \in \mathbb{R}^r \mid x_1^T \Sigma_{11}^{-1} x_1 \leq 1 \right\}$$

(h) From above we have

$$\frac{c_2}{c_1} = \sqrt{\frac{F_{\chi_2^2}^{-1}(0.9)}{F_{\chi_1^2}^{-1}(0.9)}}$$

which does not depend on Σ. This is

$$\frac{c_2}{c_1} \approx 1.3$$

3. ***The χ_n^2 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_1^2 distribution.

Solution.

One way to simplify this problem is to let $z = x/\sigma$. Then $z \sim \mathcal{N}(0, 1)$ and $y = z^2$. We also know the Gaussian pdf is

$$p^z(a) = \frac{1}{\sqrt{2\pi}} e^{-a^2/2}$$

Then since $y = z^2$ we have (from the notes)

$$p^y(b) = \frac{1}{2\sqrt{b}} \big(p^z(-\sqrt{b}) + p^z(\sqrt{b}) \big)$$

and therefore

$$p^y(b) = \frac{1}{\sqrt{2\pi b}} e^{-b/2}$$

The χ_1^2 pdf is

$$p^{\chi_1^2}(b) = \frac{1}{\Gamma(\frac{1}{2})\sqrt{2b}} e^{-b/2}$$

and so all that remains is to show that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

To do this, we have

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, dt$$

and therefore, using the change of variables $t = q^2$ we have

$$\begin{aligned}\Gamma\left(\tfrac{1}{2}\right) &= \int_0^\infty t^{-\frac{1}{2}} e^{-t}\, dt \\ &= 2\int_0^\infty e^{-q^2}\, dq\end{aligned}$$

We also know that the Gaussian pdf integrates to 1, and hence we have the famous result of Euler that

$$\int_{-\infty}^{\infty} e^{-x^2}\, dx = \sqrt{\pi}$$

Therefore $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ and we are done.