Solutions to 4F10 Pattern Processing, 2012

1. *Mixture Models and ML training*

(a) Z must satisfy

$$Z = \int \exp\left(\boldsymbol{\alpha}'\mathbf{f}(x)\right) dx$$

this ensures that the PDF integrates to 1. [10%]

(b) If $\mathbf{f}(x)$ has the form given then considering only the exponential terms

$$\alpha_1 x + \alpha_2 x^2 = -\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2}$$

Hence

$$\begin{aligned} \alpha_1 &= \frac{\mu}{\sigma^2} \\ \alpha_2 &= -\frac{1}{2\sigma^2} \end{aligned}$$

The value of Z will consist of both the standard normalisation term and the final quadratic term from the Gaussian, hence

$$Z = \frac{\sqrt{2\pi\sigma^2}}{\exp(-\mu^2/2\sigma^2)}$$

[25%]

(c)(i) The variance for each component is the same:

$$\sigma^2 = -\frac{1}{2\lambda}$$

The mean is component specific

$$\mu_m = -\frac{\alpha_m}{2\lambda}$$

Using the expression earlier

$$Z_m = \frac{\sqrt{-\pi/\lambda}}{\exp\left(\alpha_m^2/(4\lambda)\right)}$$

[15%]

(c)(ii) The auxiliary function for the general mixture model may be written as

$$\mathcal{Q}(\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}) = \sum_{m=1}^{M} \sum_{i=1}^{N} P(m|x_i, \alpha) \left[\log\left(c_m\right) - \frac{1}{2}\log(-\pi/\lambda) + \alpha_m x + \frac{\alpha_m^2}{4\lambda} + \lambda x^2 \right]$$

Differentiating with respect to α_m and equating to zero yields

$$\sum_{i=1}^{N} P(m|x_i,\alpha)\left[x_i + \frac{\hat{\alpha}_m}{2\lambda}\right] = 0$$

Yielding

$$\alpha_m = -2\lambda \frac{\sum_{i=1}^{N} P(m|x_i,\alpha)x_i}{\sum_{i=1}^{N} P(m|x_i,\alpha)}$$

[30%]

(c)(iii) Differentiating the auxiliary function with respect to λ and equating to zero yields

$$\sum_{m=1}^{M}\sum_{i=1}^{N} P(m|x_i,\alpha)\left[\frac{1}{2\lambda} - \frac{\alpha_m^2}{4\lambda^2} + x_i^2\right] = 0$$

will show that the two sets of parameters are functions of each other. This complicates the estimation process (for standard Gaussians this is not the case). Need to substitute the estimation in the previous stage so

$$\sum_{i=1}^{N} P(m|x_i,\alpha)\left[\frac{1}{2\lambda} - \left(\frac{\sum_{i=1}^{N} P(m|x_i,\alpha)x}{\sum_{i=1}^{N} P(m|x_i,\alpha)}\right)^2 + x^2\right] = 0$$

which yields

$$\frac{1}{2\hat{\lambda}} = -\frac{1}{\sum_{m=1}^{M}\sum_{i=1}^{N} P(m|x_i.\alpha)} \sum_{m=1}^{M}\sum_{i=1}^{N} P(m|x_i,\alpha)\left[x^2 - \left(\frac{\sum_{i=1}^{N} P(m|x_i,\alpha)x}{\sum_{i=1}^{N} P(m|x_i,\alpha)}\right)^2\right]$$

[20%]

Examiner's comment:

This question looked at members of the exponential family, their relationship to Gaussian distributions and mixture model training. This was a popular question, and generally well answered. However it was disappointing that so few candidates could derive the update formulae for all the model parameters.

2. *Bayes' Decision Rule and Probability of Error*

(a) Bayes' decision rule for a two class problem is

$$\text{Decide} \begin{cases} \text{Class } \omega_1 & \text{if } P(\omega_1|\boldsymbol{x}) > P(\omega_2|\boldsymbol{x}); \\ \text{Class } \omega_2 & \text{Otherwise} \end{cases}$$

[10%]

(b) For a 2-class problem the decision rule will split the observation space into two regions

- $\mathcal{R}_1$: observation classified as ω_1
- $\mathcal{R}_2$: observation classified as ω_2

$$\begin{aligned} P(\text{error}) &= P(\boldsymbol{x} \in \mathcal{R}_2, \omega_1) + P(\boldsymbol{x} \in \mathcal{R}_1, \omega_2) \\ &= P(\boldsymbol{x} \in \mathcal{R}_2|\omega_1)P(\omega_1) + P(\boldsymbol{x} \in \mathcal{R}_1|\omega_2)P(\omega_2) \\ &= \int_{\mathcal{R}_2} p(\boldsymbol{x}|\omega_1)P(\omega_1)d\boldsymbol{x} + \int_{\mathcal{R}_1} p(\boldsymbol{x}|\omega_2)P(\omega_2)d\boldsymbol{x} \end{aligned}$$

[15%]

(c)(i) A point that lies on the decision boundary satisfies

$$\log(p(\boldsymbol{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)) + \log(P(\omega_1)) = \log(p(\boldsymbol{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)) + \log(P(\omega_2))$$

Substituting in the expressions for the covariance and priors yields

$$\boldsymbol{\mu}_1'\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_1'\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2'\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_2'\boldsymbol{\mu}_2$$

Yielding the equation of a straight-line

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\,\boldsymbol{x} = \frac{1}{2}\left(\boldsymbol{\mu}_1'\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2'\boldsymbol{\mu}_2\right)$$

[25%]

(c)(ii) The decision boundary in (c)(i) defines the two regions. For this form of distribution the posterior will only be a function of the perpendicular distance from the decision boundary. Projecting the distribution for class along this line

$$\mathcal{N}\left(x; (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'\,\boldsymbol{\mu}_1/K; 1\right)$$

where $K^2 = ||\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1||^2$. Perpendicular to this direction will just integrate out to 1 for all positions.

The decision boundary is the projection of the point half way between the means onto this line. This projection point is

$$a = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)'(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)/2K$$

Considering the first element of the probability of error

$$\int_{\mathcal{R}_2} p(\boldsymbol{x}|\omega_1)P(\omega_1)d\boldsymbol{x} = \frac{1}{2}\int_a^\infty \mathcal{N}\left(x; (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'\,\boldsymbol{\mu}_1/K; 1\right) d\boldsymbol{x}$$

Offsetting the mean of the integral to 0 yields the required form

$$\int_{\mathcal{R}_2} p(\boldsymbol{x}|\omega_1)P(\omega_1)d\boldsymbol{x} = \frac{1}{2}\int_a^\infty \mathcal{N}\left(x; 0; 1\right) d\boldsymbol{x}$$

where a is now

$$\begin{aligned} a &= \frac{1}{2K}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \frac{1}{K}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'\,\boldsymbol{\mu}_1 \\ &= \frac{1}{2K}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ &= \frac{1}{2}\sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)} \end{aligned}$$

Since the probability of error for the second expression will be the same this is the value of a. [30%]

(d) In practice classifiers do not satisfy the conditions for generative models to be optimal - mention

- finite training sets
- correct form of PDFs
- priors are correct

[20%]

Examiner's comment:

This questions examined the students' knowledge of expected error rates. The question was straight-forward and was attempted by all the candidates. The main problem encountered that that rather than projecting onto the line between the two means, the average mean vector was used.

3. *Training Logistic Regression and Regularisation*

(a) The log-probability of the data from class ω_1 can be written as

$$\begin{aligned}\mathcal{L}(\mathbf{b}) &= \sum_{i=1}^{n} (y_i \log(P(\omega_1|\mathbf{x}_i, \mathbf{b})) + (1 - y_i) \log(P(\omega_2|\mathbf{x}_i, \mathbf{b})) \\ &= \sum_{i=1}^{n} (y_i \log(P(\omega_1|\mathbf{x}_i, \mathbf{b})) + (1 - y_i) \log(1 - P(\omega_1|\mathbf{x}_i, \mathbf{b}))\end{aligned}$$

This will yield a linear decision boundaries passing through the origin in the space defined by $\boldsymbol{\phi}(\mathbf{x})$.. [15%]

(a)(ii) Differentiating

$$\begin{aligned}\frac{\partial}{\partial \mathbf{b}} P(\omega_1|\mathbf{x}, \mathbf{b}) &= \frac{\exp(-\mathbf{b}'\boldsymbol{\phi}(\mathbf{x}))}{(1 + \exp(-\mathbf{b}'\boldsymbol{\phi}(\mathbf{x})))^2} \boldsymbol{\phi}(\mathbf{x}) \\ &= P(\omega_1|\mathbf{b}, \mathbf{x})(1 - P(\omega_1|\mathbf{b}, \mathbf{x}))\boldsymbol{\phi}(\mathbf{x})\end{aligned}$$

Thus

$$\begin{aligned}\frac{\partial}{\partial \mathbf{b}} \mathcal{L}(\mathbf{b}) &= \sum_{i=1}^{n} \boldsymbol{\phi}(\mathbf{x}_i) \left(y_i(1 - P(\omega_1|\mathbf{b}, \mathbf{x}_i)) - (1 - y_i)P(\omega_1|\mathbf{b}, \mathbf{x}_i)\right) \\ &= \sum_{i=1}^{n} \boldsymbol{\phi}(\mathbf{x}_i) \left(y_i - P(\omega_1|\mathbf{b}, \mathbf{x}_i)\right)\end{aligned}$$

This can be used in a gradient style approach where

$$\mathbf{b}^{(k+1)} = \mathbf{b}^{(k)} + \eta \left.\frac{\partial}{\partial \mathbf{b}} \mathcal{L}(\mathbf{b})\right|_{\mathbf{b}^{(k)}}$$

[30%]

(b)(i) The posteriors for ω_1 each of the points are given by

$$\frac{1}{1 + \exp(-\alpha)}, \frac{1}{1 + \exp(+\alpha)}$$
$$\frac{1}{1 + \exp(-\alpha)}, \frac{1}{1 + \exp(+\alpha)}$$

Two approaches can be adopted, either substitute into the likelihood, or the derivative.

Simplest is to substitute into the original likelihood yields

$$\begin{aligned}\mathcal{L}(\alpha) &= \log\left(\frac{1}{1 + \exp(-\alpha)}\right) + \log\left(\frac{1}{1 + \exp(\alpha)}\right) + \log\left(\frac{\exp(-\alpha)}{1 + \exp(-\alpha)}\right) \log\left(\frac{\exp(+\alpha)}{1 + \exp(+\alpha)}\right) \\ &= 2\log\left(\frac{1}{1 + \exp(-\alpha)}\right) + 2\log\left(\frac{1}{1 + \exp(\alpha)}\right)\end{aligned}$$

Need to maximise this expression - differentiate and equate to zero

$$2\left(\frac{\exp(-\alpha)}{1+\exp(-\alpha)}\right) - 2\left(\frac{\exp(\alpha)}{1+\exp(\alpha)}\right) = 0$$

Thus

$$\alpha = 0$$

This does not yield a reasonable classifier, all points have the same posterior. [25%]

(b)(ii) For the form of transformation specified only the second element can discriminate, s seen in the SVM lectures. Simplest solution is to set

$$\mathbf{b} = \begin{bmatrix} 0 \\ -\alpha \\ 0 \end{bmatrix}$$

The posteriors for class 1 are then for the data for class 1 and then class 2

$$\frac{1}{1+\exp(-\alpha)}, \frac{1}{1+\exp(\alpha)}$$

The points are now correctly classified. Note the requirement for $\alpha > 0$ otherwise it classifies it exactly incorrectly! [30%]

Examiner's comment:

A question based on training a logistic regression classifier and then applying a kernel on the feature-space. Though a popular question and reasonably well done, it was disappointing that more candidates did not notice that the points selected were for XOR and the feature-space chosen was described in lectures for solving this problem.

4. *Non-Linear Regression and Gaussian Processes*

(a) By inspection this form of interpolation is a non-linear function of the observations. The form given is exponential in nature which can yield a feature space of the same dimension as the number of observations. σ^2 determines how smooth the regression will be. [20%]

(b) The regression process can be written as

$$\boldsymbol{y} = \boldsymbol{\Phi}\mathbf{w} + \sigma_{\mathrm{n}}^2\mathbf{I}$$

By inspection the mean is zero as both $\mathbf{w}$ and the noise are zero mean. Following the notation from lectures

$$\boldsymbol{\Sigma}_y = \sigma_{\mathrm{w}}^2\boldsymbol{\Phi}\boldsymbol{\Phi}' + \sigma_{\mathrm{n}}^2\mathbf{I}$$

where

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi_1(\boldsymbol{x}_1) & \dots & \phi_H(\boldsymbol{x}_1) \\ \vdots & \ddots & \vdots \\ \phi_1(\boldsymbol{x}_n) & \dots & \phi_H(\boldsymbol{x}_n) \end{bmatrix}$$

Expanding out this representation yields for a particular element

$$[\boldsymbol{\Sigma}_{\mathrm{y}}]_{ij} = \sigma_{\mathrm{w}}^2 \sum_{h=1}^{H} \exp\left(-\frac{(x_i-\mu_h)^2}{2\sigma^2}\right)\exp\left(-\frac{(x_j-\mu_h)^2}{2\sigma^2}\right) + \sigma_{\mathrm{n}}^2\delta(i-j)$$

[35%]

(c) In the limit as $H \to \infty$ yields (grouping elements in μ_h and μ_h^2 together)

$$\begin{aligned}
[\boldsymbol{\Sigma}_{\mathrm{y}}]_{ij} &= \sigma_{\mathrm{h}}^2 \int \exp\left(-\frac{(x_i-\mu_h)^2}{2\sigma^2}\right)\exp\left(-\frac{(x_j-\mu_h)^2}{2\sigma^2}\right) p(\mu_h)d\mu_h + \sigma_{\mathrm{n}}^2\delta(i-j) \\
&= \frac{\sigma_{\mathrm{h}}^2}{\sqrt{2\pi\sigma_{\mathrm{h}}^2}} \int \exp\left(-\frac{2\sigma_h^2+\sigma^2}{2\sigma^2\sigma_{\mathrm{h}}^2}\mu_h^2 + 2\frac{(x_i+x_j)}{2\sigma^2}\mu_h - \frac{(x_i^2+x_j^2)}{2\sigma^2}\right) d\mu_h + \sigma_{\mathrm{n}}^2\delta(i-j)
\end{aligned}$$

Noting that $\sigma_{\mathrm{h}}^2 \gg \sigma^2$ (the prior has no impact) this can be approximated by

$$[\boldsymbol{\Sigma}_{\mathrm{y}}]_{ij} \approx \frac{\sigma_{\mathrm{h}}^2}{\sqrt{2\pi\sigma_{\mathrm{h}}^2}} \int \exp\left(-\frac{1}{\sigma^2}\mu_h^2 + 2\frac{(x_i+x_j)}{2\sigma^2}\mu_h - \frac{(x_i^2+x_j^2)}{2\sigma^2}\right) d\mu_h + \sigma_{\mathrm{n}}^2\delta(i-j)$$

Using the equality this can be expressed as

$$\begin{aligned}
[\boldsymbol{\Sigma}_{\mathrm{y}}]_{ij} &\approx \frac{2\pi\sigma^2\sigma_{\mathrm{h}}^2}{\sqrt{2\pi\sigma_{\mathrm{h}}^2}} \int \mathcal{N}(\mu_h; x_i, \sigma^2)\mathcal{N}(\mu_h; x_j, \sigma^2)d\mu_h + \sigma_{\mathrm{n}}^2\delta(i-j) \\
&= \frac{2\pi\sigma^2\sigma_{\mathrm{h}}^2}{\sqrt{2\pi\sigma_{\mathrm{h}}^2}} \frac{1}{2\sqrt{\pi\sigma^2}} \exp\left(-\frac{1}{4\sigma^2}(x_i-x_j)^2\right) + \sigma_{\mathrm{n}}^2\delta(i-j) \\
&= \sqrt{\frac{\sigma^2\sigma_{\mathrm{h}}^2}{2}} \exp\left(-\frac{1}{4\sigma^2}(x_1-x_2)^2\right) + \sigma_{\mathrm{n}}^2\delta(i-j)
\end{aligned}$$

The alternative approach is to compute this directly

$$
\begin{aligned}
[\Sigma_{\mathrm{y}}]_{ij} &\approx \frac{\sigma_{\mathrm{h}}^2}{\sqrt{2\pi\sigma_{\mathrm{h}}^2}} \int \exp\left(\frac{2}{2\sigma^2}\mu_h^2 + 2\frac{(x_i+x_j)}{2}\frac{1}{\sigma^2}\mu_h - \frac{(x_i^2+x_j^2)}{2}\frac{1}{\sigma^2}\right) d\mu_h + \sigma_{\mathrm{n}}^2\delta(i-j) \\
&= \frac{\sigma_{\mathrm{h}}^2\sqrt{\pi\sigma^2}}{\sqrt{2\pi\sigma_{\mathrm{h}}^2}} \exp\left(\frac{(x_i+x_j)^2}{4\sigma^2} - \frac{(x_i^2+x_j^2)}{2}\frac{1}{\sigma^2}\right) + \sigma_{\mathrm{n}}^2\delta(i-j) \\
&= \sqrt{\frac{\sigma^2\sigma_{\mathrm{h}}^2}{2}} \exp\left(-\frac{1}{4\sigma^2}(x_1-x_2)^2\right) + \sigma_{\mathrm{n}}^2\delta(i-j)
\end{aligned}
$$

[35%]

(d) Though the number of weights increases to infinity, the regression process becomes a function of the model parameters. Thus the effective number of free parameters is determined by the number of observations. [10%]

Examiner's comment:

Those candidates that attempted this question did reasonably well. The question was based on Gaussian processes. As in previous years this topic was unpopular with the students.

5. *Classification and Regression Trees*

(a)(i) The general attributes that should be satisfied by a node impurity function, $\phi()$, are

- $\phi()$ is a maximum when $P(\omega_i) = 1/K$ for all i
- $\phi()$ is at a minimum when $P(\omega_i) = 1$ and $P(\omega_j) = 0, j \neq i$.
- It is symmetric function (i.e. the order of the class probabilities doesn't matter). [20%]

(a)(ii) $P(\omega_i)$ should be calculated as the fraction of the observations belonging to class ω_i associated with that node.

The Gini impurity measure may be written as

$$\begin{aligned}\sum_{i \neq j} P(\omega_i)P(\omega_j) &= \sum_i P(\omega_i) \sum_{j \neq i} P(\omega_j) \\ &= \sum_i^K P(\omega_i)(1 - P(\omega_i)) \\ &= 1 - \sum_{i=1}^K (P(\omega_i))^2\end{aligned}$$

This function satisfies all the attributes. Using the second form of the Gini impurity measure

- The function is a maximum when all values are equal (various acceptable proof including

  $$(x + \delta)^2 + (x - \delta)^2 = 2x^2 + \delta^2$$

  This is greater than $2x^2$ so the impurity measure will be less.)
- The value is at a minimum when $P(\omega_i) = 1$ (all other classes zero).
- By inspection it will be symmetric from the second form. [25%]

(b)(i) The two possible splits are for attribute one and attribute 2

- Attribute 1: 1 2/5 ω_1, 0 2/3 ω_1
- Attribute 2: 1 1/5 ω_1, 0 3/3 ω_1

Attribute 2 is clearly better as both splits are more pure. The actual calculation of the change in impurity function is

$$1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 - \frac{5}{8}\left(1 - \left(\frac{1}{5}\right)^2 - \left(\frac{4}{5}\right)^2\right) - \frac{3}{8}\left(1 - \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2\right) = \frac{1}{2} - \frac{1}{5} = \frac{3}{10}$$

[20%]

(b)(ii) Two of the feature vectors are identical. These cannot be split using this data. All other symbols may be perfectly classified. Thus the lowest impurity measure is given by

$$\frac{1}{4}\left(1-\left(\frac{1}{2}\right)^2-\left(\frac{1}{2}\right)^2\right)=\frac{1}{8}$$

as all other nodes will be correct. [15%]

(c) Since misclassification costs are now considered the appropriate criterion would be the misclassification criterion

$$1-\max_i\{P(\omega_i)\}$$

The cost function would then be altered so that the cost is twice this when the maximum is class ω_1. [20%]

Examiner's comment:

This question examined the candidates knowledge of decision trees and the Gini impurity function. A very straightforward question and very well done by the candidates that attempted the question. This was the least popular question on the paper.