

Paper 4F10: Statistical Pattern Processing
 STATISTICAL PATTERN RECOGNITION

Examples Paper 2

Straightforward questions are marked †

*Tripos standard (but not necessarily Tripos length) questions are marked **

Support Vector Machines

1. † A binary classifier is to be trained. What are the limitations of linear decision classifiers and why do non-linear mappings of the feature space allow improved discrimination? Under what conditions is it guaranteed that a non-linear mapping will allow perfect classification of the data?
2. For the XOR problem described in lecture notes show that the solution given satisfies the training conditions given. What is the equation of the final decision boundary?
3. The following data is to be used for training an SVM

$$\begin{array}{l} \omega_1 : \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \begin{bmatrix} 2 \\ 2 \end{bmatrix} \quad \begin{bmatrix} 2 \\ 0 \end{bmatrix} \\ \omega_2 : \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \begin{bmatrix} 0 \\ 1 \end{bmatrix} \end{array}$$

- (a) Plot the training points and, by inspection, determine the position of the optimal, maximum margin, decision boundary.
- (b) What are the support vectors?
- (c) Express the decision boundary in terms of the Lagrange multipliers, α_i and show that this satisfies the KKT conditions.

Gaussian Processes and Relevance Vector Machines

4. † Show that using

$$P(y = +1 | \mathbf{x}, \hat{\mathbf{w}}) = \begin{cases} 1, & \hat{\mathbf{w}}' \mathbf{x} + \epsilon \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

when $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$ has the form

$$P(y = +1 | \mathbf{x}, \hat{\mathbf{w}}) = 1 - \int_{-\infty}^a \mathcal{N}(z; 0, 1) dz$$

What is the value of a ?

5. Linear regression of the form

$$y = \mathbf{w}'\mathbf{x} + \epsilon$$

where $\epsilon \sim \mathcal{N}(0; \sigma_{\mathbf{n}}^2)$ and the weights have a prior distribution of the form $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathbf{w}}^2)$ is to be used. Show that the marginal likelihood, $p(\mathbf{y}|\mathbf{X})$, is given by

$$p(\mathbf{y}|\mathbf{X}) = \mathcal{N}(\mathbf{y}; \mathbf{0}, \sigma_{\mathbf{w}}^2 \mathbf{X}'\mathbf{X} + \sigma_{\mathbf{n}}^2 \mathbf{I})$$

where the n training data samples $\{\mathbf{x}_1, y_1\}, \dots, \{\mathbf{x}_n, y_n\}$ are expressed as

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \dots & \mathbf{x}_n \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

The following equality may be used

$$(\mathbf{ABC} + \mathbf{D})^{-1} = \mathbf{D}^{-1} - \mathbf{D}^{-1}\mathbf{A}(\mathbf{CD}^{-1}\mathbf{A} + \mathbf{B}^{-1})^{-1}\mathbf{CD}^{-1}$$

6. * A Gaussian process is to be used for prediction. There are n training examples consisting of the observations, $\mathbf{x}_1, \dots, \mathbf{x}_n$ and associated targets y_1, \dots, y_n . The prediction at the point $\tilde{\mathbf{x}}$ is required. The model used for prediction is of the form

$$y = f(\mathbf{x}) + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, \sigma_{\mathbf{n}}^2)$

- Derive an expression for the distribution of the prediction. Only the systematic prediction variance should be considered.
- Rather than using all n training examples, only the first $n - 1$ are used. If the variance of the prediction using all n examples is denoted as $\text{var}_n(f(\tilde{\mathbf{x}}))$, show that

$$\text{var}_n(f(\tilde{\mathbf{x}})) \leq \text{var}_{n-1}(f(\tilde{\mathbf{x}}))$$

where $\text{var}_{n-1}(f(\tilde{\mathbf{x}}))$ is the prediction using the first $n - 1$ points.

The following matrix equality may be used

$$\begin{bmatrix} \mathbf{A} & \mathbf{b} \\ \mathbf{b}' & c \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{b}'(c - \mathbf{b}'\mathbf{A}^{-1}\mathbf{b})^{-1}\mathbf{b}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{b}(c - \mathbf{b}'\mathbf{A}^{-1}\mathbf{b})^{-1} \\ - (c - \mathbf{b}'\mathbf{A}^{-1}\mathbf{b})^{-1}\mathbf{b}'\mathbf{A}^{-1} & (c - \mathbf{b}'\mathbf{A}^{-1}\mathbf{b})^{-1} \end{bmatrix}$$

[This question illustrates that the prediction variance cannot get worse as the number of training samples increases]

7. The mean of the prediction distribution of the output y with \mathbf{x} for a relevance vector machine using basis function $\phi(\mathbf{x})$ is given by

$$\mu_y = \phi(\mathbf{x})' \boldsymbol{\mu}_w = \frac{1}{\sigma_n^2} \phi(\mathbf{x})' \left(\frac{1}{\sigma_n^2} \boldsymbol{\Phi}' \boldsymbol{\Phi} + \boldsymbol{\Lambda} \right)^{-1} \boldsymbol{\Phi}' \mathbf{y}$$

where \mathbf{y} are the training data “targets” for the training data points $\mathbf{x}_1, \dots, \mathbf{x}_n$, $\boldsymbol{\Lambda}$ is the diagonal inverse covariance matrix associated with the weights,

$$p(\mathbf{w} | \boldsymbol{\Lambda}) = \prod_{i=1}^n \mathcal{N}(w_i; 0, \lambda_i^{-1})$$

and

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi(\|\mathbf{x}_1 - \mathbf{x}_1\|) & \dots & \phi(\|\mathbf{x}_1 - \mathbf{x}_n\|) \\ \vdots & \ddots & \vdots \\ \phi(\|\mathbf{x}_n - \mathbf{x}_1\|) & \dots & \phi(\|\mathbf{x}_n - \mathbf{x}_n\|) \end{bmatrix} = \begin{bmatrix} \phi(\mathbf{x}_1)' \\ \vdots \\ \phi(\mathbf{x}_n)' \end{bmatrix}$$

Show that as $\lambda_i \rightarrow \infty$ the mean of the posterior distribution for weight w_i given the training data is zero and the variance goes to zero. Note the matrix equality in question (6) may be used.

Classification and Regression Trees

8. A tree classifier is to be built for a one-dimensional two category problem. A large number of training samples are available. These samples are drawn from two classes with equal priors. The class-conditional probability distributions for the two classes are Gaussian distributed with

$$\begin{aligned} p(x|\omega_1) &= \mathcal{N}(x; 0, 1) \\ p(x|\omega_2) &= \mathcal{N}(x; 1, 1) \end{aligned}$$

All nodes will have decisions of the form “Is $x \leq x_s$ ” where x_s is some threshold. At the top the level the value of the split threshold is x_1 . The size of the tree is limited. It is a binary tree with a root node and two non-terminal nodes yielding a total of four leaf nodes. The binary split cost is given by

$$\Delta \mathcal{I}(N) = \mathcal{I}(N) - f_L \mathcal{I}(N_L) - (1 - f_L) \mathcal{I}(N_R)$$

where f_L is the fraction of the data from the current node assigned to the left descendant. The entropy cost function is to be used. For the non-terminal node that satisfies the root node question find an expression, in terms of the cumulative density function for a Gaussian, for the binary split cost.

Non-Parameteric Techniques

9. * n samples are drawn from a Gaussian distribution with mean, μ , and variance, σ^2 . Consider a Gaussian window function of the form

$$\phi(x) = \mathcal{N}(x; 0, 1)$$

Show that the Parzen window estimate of the true distribution, $p(x) = \mathcal{N}(x, \mu, \sigma^2)$,

$$\tilde{p}(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h_n} \phi\left(\frac{x - x_i}{h_n}\right)$$

has the following properties (for small h_n):

- (a) $\mathcal{E}\{\tilde{p}(x)\} = \mathcal{N}(x; \mu, \sigma^2 + h_n^2)$.
- (b) $\text{var}[\tilde{p}(x)] \approx \frac{1}{2nh_n\sqrt{\pi}}p(x)$
- (c) $p(x) - \mathcal{E}\{\tilde{p}(x)\} \approx \frac{1}{2} \left(\frac{h_n}{\sigma}\right)^2 \left(1 - \left(\frac{x-\mu}{\sigma}\right)^2\right) p(x)$

Note the following equality may be used

$$\int_{-\infty}^{\infty} \mathcal{N}(x; v, \sigma_1^2) \mathcal{N}(v, \mu, \sigma_2^2) dv = \mathcal{N}(x, \mu, \sigma_1^2 + \sigma_2^2)$$

Speaker verification

10. * A Support Vector Machine (SVM) is to be used for speaker verification. A 1-dimensional feature-vector is used to represent each frame of data. The feature-space to be used for with the SVM with observations $\mathbf{X}_{1:T} = \{x_1, \dots, x_T\}$ is defined as

$$\Phi(\mathbf{X}_{1:T}) = \begin{bmatrix} \frac{\partial}{\partial \mu_1} \log(p(\mathbf{X}_{1:T})) \\ \vdots \\ \frac{\partial}{\partial \mu_M} \log(p(\mathbf{X}_{1:T})) \\ \frac{\partial^2}{\partial \mu_1^2} \log(p(\mathbf{X}_{1:T})) \\ \vdots \\ \frac{\partial^2}{\partial \mu_1 \partial \mu_M} \log(p(\mathbf{X}_{1:T})) \\ \vdots \\ \frac{\partial^2}{\partial \mu_M^2} \log(p(\mathbf{X}_{1:T})) \end{bmatrix}$$

where the generative model is an M -component Gaussian Mixture Model (GMM), so

$$p(\mathbf{X}_{1:T}) = \prod_{t=1}^T \sum_{m=1}^M c_m \mathcal{N}(x_t; \mu_m, \sigma_m^2)$$

- (a) Why is this form of feature-space suitable for use with SVMs when classifying variable-length data-sequences, such as in speaker verification? Why is an SVM a suitable form of classifier as M (the number of components) gets large? What is the dimensionality of the feature-space in this case?
- (b) Derive an expression for $\frac{\partial}{\partial \mu_i} \log(p(\mathbf{X}_{1:T}))$. This should be expressed in terms of the $P(i|x_t)$, the posterior probability that component i generated the observation.
- (c) Hence show that

$$\frac{\partial^2}{\partial \mu_j \partial \mu_i} \log(p(\mathbf{X}_{1:T})) = - \sum_{t=1}^T P(i|x_t) P(j|x_t) \frac{(x_t - \mu_j)(x_t - \mu_i)}{\sigma_i^2 \sigma_j^2}$$

Do you expect these second-order derivative terms to help in classification?