Hong Kong University of Science and Technology

COMP527: Pattern Recognition

Fall 2001

Solution for Midterm Examination

1. The statement is incorrect.

   The objective of classifier training is to minimize the classification error of the test set, which is a sufficiently large random sample drawn from the feature space according to the underlying data distribution. Minimizing the classification error of the training set or the validation set is only an indirect way of trying to achieve the ultimate objective.

2. The statement is correct.

   Bayesian estimation integrates the prior parameter distribution $p(\theta)$ and the likelihood $p(D|\theta)$ of data set D to obtain the posterior parameter distribution $p(\theta|D)$, from which MAP estimation is performed to find a point estimate of the parameter θ that maximizes $p(\theta|D)$. ML estimation, on the other hand, finds a point estimate of the parameter θ that maximizes the likelihood $p(D|\theta)$. In other words, we have:

$$\begin{aligned}\theta_{\text{MAP}} &= \arg\max_{\theta} p(\theta|D) = \arg\max_{\theta} \frac{p(D|\theta)p(\theta)}{p(D)} = \arg\max_{\theta} p(D|\theta)p(\theta)\\ \theta_{\text{ML}} &= \arg\max_{\theta} p(D|\theta)\end{aligned}$$

3. (a) The conditional risks are:

$$\begin{aligned}R(\alpha_1|\mathbf{x}) &= \lambda_{11}P(\omega_1|\mathbf{x}) + \lambda_{12}P(\omega_2|\mathbf{x}) + \lambda_{13}P(\omega_3|\mathbf{x})\\ &= P(\omega_2|\mathbf{x}) + P(\omega_3|\mathbf{x})\\ R(\alpha_2|\mathbf{x}) &= \lambda_{21}P(\omega_1|\mathbf{x}) + \lambda_{22}P(\omega_2|\mathbf{x}) + \lambda_{23}P(\omega_3|\mathbf{x})\\ &= 2P(\omega_1|\mathbf{x}) + 2P(\omega_3|\mathbf{x})\\ R(\alpha_3|\mathbf{x}) &= \lambda_{31}P(\omega_1|\mathbf{x}) + \lambda_{32}P(\omega_2|\mathbf{x}) + \lambda_{33}P(\omega_3|\mathbf{x})\\ &= 4P(\omega_1|\mathbf{x}) + 4P(\omega_2|\mathbf{x})\end{aligned}$$

   (b) The minimum-risk classification rule is:

$$\text{Decide}\begin{cases}\omega_1 & P(\omega_2|\mathbf{x}) < 2P(\omega_1|\mathbf{x}) + P(\omega_3|\mathbf{x}) \text{ and } P(\omega_3|\mathbf{x}) < 4P(\omega_1|\mathbf{x}) + 3P(\omega_2|\mathbf{x})\\ \omega_2 & 2P(\omega_1|\mathbf{x}) + P(\omega_3|\mathbf{x}) < P(\omega_2|\mathbf{x}) \text{ and } 2P(\omega_3|\mathbf{x}) < 2P(\omega_1|\mathbf{x}) + 4P(\omega_2|\mathbf{x})\\ \omega_3 & 4P(\omega_1|\mathbf{x}) + 3P(\omega_2|\mathbf{x}) < P(\omega_3|\mathbf{x}) \text{ and } 2P(\omega_1|\mathbf{x}) + 4P(\omega_2|\mathbf{x}) < 2P(\omega_3|\mathbf{x})\\ \text{reject} & \text{otherwise}\end{cases}$$

4. (a) See page 4.37 of notes.

   (b) Using typical nonparametric density estimation methods such as Parzen-window estimation, the likelihood of a test example is estimated based only on training examples in the vicinity of the test example in the feature space. When only limited data are available for training, the estimate obtained can be very inaccurate. On the other hand, parametric methods can estimate the density function well even with limited data, as long as the assumption about the parametric function form is valid.

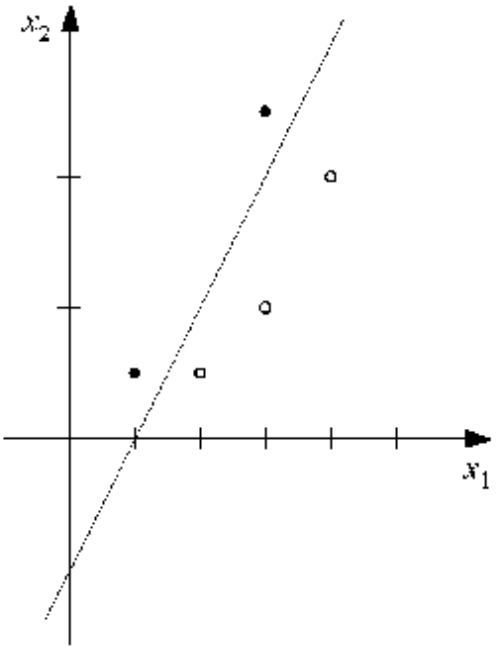
5. Since this is a 2-dimensional problem, we can plot the data points and draw a straight line to separate the training examples into two classes.

   Let the equation of the straight line be $x_1 + ax_2 + b = 0$. Since it passes through the points $(1, 0)$ and $(0, -2)$, the following two equations should hold:

$$\begin{cases} 1 + b = 0 \\ -2a + b = 0 \end{cases}$$

   Solving these two equations, we can see that $a = -1/2$ and $b = -1$. Thus the equation of the straight line can be expressed as $2x_1 - x_2 - 2 = 0$, which implies that the linear discriminant function $g(\mathbf{x}) = 2x_1 - x_2 - 2$ can solve the classification task. To verify this, we substitute the five points into $g(\mathbf{x})$:

$$\begin{aligned} g(2,1) &= 1 \quad (> 0) \\ g(3,2) &= 2 \quad (> 0) \\ g(4,4) &= 2 \quad (> 0) \\ g(1,1) &= -1 \quad (< 0) \\ g(3,5) &= -1 \quad (< 0) \end{aligned}$$



6. The first part of the derivation is independent of the type of output units used. The total error is given by:

$$E = \sum_{p=1}^{n} E^{(p)} \quad \text{where} \quad E^{(p)} = \frac{1}{2} \sum_{k=1}^{c} (t_k^{(p)} - y_k^{(p)})^2$$

   The gradient term with respect to a hidden-to-output weight is:

$$\frac{\partial E}{\partial w_{kj}} = \sum_{p=1}^{n} \frac{\partial E^{(p)}}{\partial w_{kj}} = \sum_{p=1}^{n} \frac{\partial E^{(p)}}{\partial y_k^{(p)}} \frac{\partial y_k^{(p)}}{\partial x_k^{(p)}} \frac{\partial x_k^{(p)}}{\partial w_{kj}} = -\sum_{p=1}^{n} (t_k^{(p)} - y_k^{(p)}) f_k'(x_k^{(p)}) y_j^{(p)}$$

The activation function of the softmax output units is:

$$f_k(x_k) = y_k = \frac{\exp(x_k)}{\sum_{k'=1}^{c} \exp(x_{k'})} \quad \text{where} \quad x_k = \sum_{j=0}^{h} w_{kj} y_j$$

The first derivative of the activation function is:

$$\begin{aligned} f'_k(x_k) = \frac{dy_k}{dx_k} &= \frac{(\sum_{k'=1}^{c} \exp(x_{k'}))e^{x_k} - e^{x_k} e^{x_k}}{(\sum_{k'=1}^{c} \exp(x_{k'}))^2} \\ &= \frac{\sum_{k'=1}^{c} \exp(x_{k'}) - e^{x_k}}{\sum_{k'=1}^{c} \exp(x_{k'})} \frac{e^{x_k}}{\sum_{k'=1}^{c} \exp(x_{k'})} \\ &= (1 - y_k) y_k \end{aligned}$$

Thus, the weight change is:

$$\Delta w_{kj} = -\eta \frac{\partial E}{\partial w_{kj}} = \eta \sum_{p=1}^{n} \delta_k^{(p)} y_j^{(p)} \quad \text{where} \quad \delta_k^{(p)} = (t_k^{(p)} - y_k^{(p)})(1 - y_k^{(p)}) y_k^{(p)}$$