E0 270: MACHINE LEARNING (JAN-APRIL 2024)

PROBLEM SHEET #1
INDIAN INSTITUTE OF SCIENCE

1. Suppose we have two features $x = (x_1, x_2)$ and the two class-conditional densities, $P(x|\omega = 1)$ and $P(x|\omega = 2)$, are 2D Gaussians distributions centered at points $(4, 11)$ and $(10, 3)$ respectively with same covariance $\Sigma = 3I$ (where I is the identity matrix). Suppose the priors are $P(\omega = 1) = 0.6$ and $P(\omega = 2) = 0.4$. Using bayes rule find the two discriminant functions $g_1(x)$ and $g_2(x)$? Derive the equation for decision boundary?

   Solution:
   Discriminant functions are given by

$$\begin{aligned} g_i(x) &= \log P(x|w_i) + \log P(w_i) \\ &= \log \frac{1}{\sqrt{2\pi.3}} - \frac{1}{2}||x - \mu_i||^2 \frac{1}{3} + \log P(w_i). \end{aligned}$$

   Substitute μ_i and $P(w_i)$ to obtain the discriminant functions.
   For obtaining the decision boundary, set $g_1(x) = g_2(x)$.

$$-\frac{1}{6}||x - (4, 11)||^2 + \log 0.6 = -\frac{1}{6}||x - (10, 3)||^2 + \log 0.4.$$

   Simplifying the above gives the equation for the decision boundary.

2. In a two class, two dimensional classification task the feature vectors are generated by two normal distributions sharing the same covariance matrix:

$$\Sigma = \begin{bmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{bmatrix}, \ \Sigma^{-1} = \begin{bmatrix} 0.95 & -0.15 \\ -0.15 & 0.55 \end{bmatrix}, \ |\Sigma| = 2$$

   and the mean vectors $\mu_1 = [0, 0]^T$ and $\mu_2 = [3, 3]^T$ respectively. Classify the vector $[1.0, 2.2]^T$ according to bayes classifier? (assume uniform prior) (from https://www.cse.unr.edu/ bebis/CS479/Handouts/)

   Solution:
   Since the prior is uniform, we have $P(w_1) = P(w_2)$, so the Bayes classifier just classifies based on the likelihood. Since Σ is the same for both classes,

the discriminant functions can be written as only the second term of the log likelihood, as:

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T\Sigma^{-1}(x-\mu_i).$$

For classifying the vector $(1.0, 2.2)$, just substitute it as x, and also substitute the values of μ_i and Σ in the above discriminant functions and choose the one with the higher value.

3. Consider a linear model of the form

$$y(x,w) = w_0 + \sum_i^D w_i x_i$$

together with a sum-of-squares error function of the form

$$E_D(w) = 0.5 * \sum_{n=1}^N [y(x_n,w) - t_n]^2$$

Now suppose that Gaussian noise η_i with zero mean and variance σ^2 is added independently to each of the variables x_i. By making use of $\mathcal{E}[\eta_i] = 0$ and $\mathcal{E}[\eta_i\eta_j] = \delta_{ij}\sigma^2$, show that minimizing E_D averaged over the noise distribution is equivalent to minimizing the sum-of-squares error for noise-free input variables with the addition of a weight-decay regularization term, in which the bias parameter w_0 is omitted from the regularizer. (Bishop 3.4)

Solution:

Since noise is added to each element of x, we have $\tilde{x}_{ni} = x_{ni} + \eta_{ni}$. So, the linear model becomes

$$\begin{aligned} y(\tilde{x}_n, w) &= w_0 + \sum_{i=1}^D w_i \tilde{x}_{ni} \\ &= w_0 + \sum_{i=1}^D w_i (x_{ni} + \eta_{ni}) \end{aligned}$$

The sum-of-squares error function becomes

$$
\begin{aligned}
\tilde{E}_D(w) &= \frac{1}{2N}\sum_{n=1}^{N}\left[w_0 + \sum_{i=1}^{D} w_i(x_{ni}+\eta_{ni}) - t_n\right]^2 \\
&= \frac{1}{2N}\sum_{n=1}^{N}\left[w_0 + \sum_{i=1}^{D} w_i x_{ni} - t_n + \sum_{i=1}^{D} w_i\eta_{ni}\right]^2 \\
&= \frac{1}{2N}\sum_{n=1}^{N}\left[w_0 + \sum_{i=1}^{D} w_i x_{ni} - t_n\right]^2 + \frac{1}{2N}\sum_{n=1}^{N}\left(\sum_{i=1}^{D} w_i\eta_{ni}\right)^2 \\
&\qquad + \frac{1}{2N}\sum_{n=1}^{N}\left(\sum_{i=1}^{D}\eta_{ni}\right)\left(w_0 + \sum_{i=1}^{D} w_i x_{ni} - t_n\right) \\
&= \frac{1}{2N}\sum_{n=1}^{N}\left[w_0 + \sum_{i=1}^{D} w_i x_{ni} - t_n\right]^2 + \frac{1}{2N}\sum_{n=1}^{N}\left(\sum_{i=1}^{D} w_i^2\eta_{ni}^2 + \sum_{i\neq j} w_i w_j \eta_{ni}\eta_{nj}\right) \\
&\qquad + \frac{1}{2N}\sum_{n=1}^{N}\left(\sum_{i=1}^{D}\eta_{ni}\right)\left(w_0 + \sum_{i=1}^{D} w_i x_{ni} - t_n\right)
\end{aligned}
$$

Taking expectation wrt η_{ni}'s, since the noise terms are zero mean and independent, only the first and second terms remain while the other go to zero. So, we have

$$\tilde{E}_D(w) = E_D(w) + \frac{\sigma^2}{2}\sum_{i=1}^{D} w_i^2.$$

4. A student needs to achieve a decision on which courses to take, based only on his first lecture. From previous experience he knows the following

Quality of Course	Good	Fair	Bad
Probability ($P(\omega_j)$)	0.2	0.4	0.4

These are the priors. The student also knows the class conditionals

$P(x\|\omega_j)$	Good	Fair	Bad
Interesting Lecture	0.8	0.5	0.1
Boring Lecture	0.2	0.5	0.9

He also knows the loss function for the actions

$\lambda(a_i\|\omega_j)$	Good	Fair	Bad
Taking the course	0	5	10
Not taking the course	20	5	0

What is the optimal decision by minimizing the risk if he found the lecture for a course interesting? ($http://www.cs.haifa.ac.il/~rita/ml\_course$)

Solution:

$$R(\alpha|x = interesting) = \sum_{i \in \{\text{good, fair, bad}\}} \lambda(\alpha|w_i)p(w_i|x),$$

where $\alpha \in \{\text{take, not take}\}$. Need to find $p(w_i|x)$ for each w_i as

$$\begin{aligned} p(w_i|x) &= \frac{p(x|w_i)p(w_i)}{\sum_j p(x|w_j)p(w_j)} \\ &= \frac{p(x|w_i)p(w_i)}{0.2 * 0.8 + 0.4 * 0.5 + 0.4 * 0.1}. \end{aligned}$$

5. In many pattern classification problems one has the option either to assign to one of c *classes*, or to *reject* it as being unrecognizable. If the cost for rejects is not too high, rejection may be a desirable action. Let

$$\begin{aligned} \lambda(\alpha_i|\omega_j) &= 0 \; if \;\; i = j \;\&\; i, j = 1, ..., c \\ &= \lambda_r \;\; i = c + 1 \\ &= \lambda_s \;\; otherwise \end{aligned}$$

where λ_r is the loss incurred for choosing the $(c + 1)$th action, rejection, and λ_s is the loss incurred for making a substitution error. Show that the minimum risk is obtained if we decide ω_i if $P(\omega_i|x) \geq P(\omega_j|x)$ for all j and if $P(\omega_i|x) \geq 1 - \lambda_r/\lambda_s$, and reject otherwise. What happens if $\lambda_r = 0$? What happens if $\lambda_r > \lambda_s$?

Solution:

The conditional risk is minimized by choosing action α that minimizes

$$\sum_{j=1}^{c} \lambda(\alpha|w_j)P(w_j|x).$$

Now, for actions α_i, $i = 1, \ldots, c$, the above expression becomes

$$\begin{aligned} \sum_j \lambda(\alpha_i|w_j)P(w_j|x) = \sum_{j \neq i} \lambda_s P(w_j|x) &= \lambda_s \sum_{j \neq i} P(w_j|x) \\ &= \lambda_s \left(1 - P(w_i|x)\right). \end{aligned}$$

The above expression is minimized for $i = \underset{i}{\operatorname{argmax}} P(w_i|x)$, in which case it becomes $\lambda_s (1 - P(w_{max}|x))$. Now, the conditional risk of rejecting is

$$\sum_j \lambda(\alpha_{c+1}|w_j)P(w_j|x) = \sum_j \lambda_r P(w_j|x) = \lambda_r.$$

Therefore, the chosen action will be rejection if and only if its conditional risk is lesser, i.e,

$$\begin{aligned} \lambda_r &< \lambda_s (1 - P(w_{max}|x)), \\ \text{i.e, } \frac{\lambda_r}{\lambda_s} &< 1 - P(w_{max}|x) \\ \text{i.e, } P(w_{max}|x) &< 1 - \frac{\lambda_r}{\lambda_s}. \end{aligned}$$

6. Let the conditional densities for a two-category one-dimensional problem be given by the Cauchy distribution (Duda Hart Prob 7 & 8)

$$P(x|\omega_i) = \frac{1}{\pi b}\frac{1}{1 + (\frac{x-a_i}{b})^2} \text{ for } i = 1, 2.$$

(a) Find the minimum error decision boundary for 0-1 loss assuming uniform prior ?
(b) Show the the minimum probability of error is given by

$$P(error) = \frac{1}{2} - \frac{1}{\pi}\tan^{-1}|\frac{a_1 - a_2}{2b}|.$$

Solution:
(a) Let $a_1 < a_2$. Since the prior probabilities of both classes are the same, the decision boundary corresponds to the region where the class conditional likelihoods of both classes are the same, i.e,

$$\begin{aligned} P(x|w_1) &= P(x|w_2) \\ \text{i.e, } \frac{1}{\pi b}\frac{1}{1 + \left(\frac{x-a_1}{b}\right)^2} &= \frac{1}{\pi b}\frac{1}{1 + \left(\frac{x-a_2}{b}\right)^2} \\ \text{i.e, } (x - a_1)^2 &= (x - a_2)^2 \\ \text{i.e, } |x - a_1| &= |x - a_2| \\ \text{i.e, } x &= \frac{a_1 + a_2}{2}. \end{aligned}$$

(b) Let R_1 be the region where w_1 is predicted and R_2 be the region where w_2 is predicted. Then

$$\begin{aligned}
P(error) &= \int_x p(error, x)dx \\
&= \int_{R_2} p(x|w_1)p(w_1)dx + \int_{R_1} p(x|w_2)p(w_2)dx \\
&= \int_{\frac{a_1+a_2}{2}}^{\infty} \frac{1}{\pi b}\frac{1}{1+\left(\frac{x-a_1}{b}\right)^2}\frac{1}{2}dx + \int_{-\infty}^{\frac{a_1+a_2}{2}} \frac{1}{\pi b}\frac{1}{1+\left(\frac{x-a_2}{b}\right)^2}\frac{1}{2}dx \\
&= \frac{1}{2\pi b}\int_{\frac{a_2-a_1}{2b}}^{\infty} \frac{b}{1+y^2}dy + \frac{1}{2\pi b}\int_{-\infty}^{\frac{a_1-a_2}{2b}} \frac{b}{1+y^2}dy \\
&= \frac{1}{2\pi}\tan^{-1} y\Big|_{\frac{a_2-a_1}{2b}}^{\infty} + \frac{1}{2\pi}\tan^{-1} y\Big|_{-\infty}^{\frac{a_1-a_2}{2b}} \\
&= \frac{1}{2\pi}\left[\frac{\pi}{2} - \tan^{-1}\left(\frac{a_2-a_1}{2b}\right) + \tan^{-1}\left(\frac{a_1-a_2}{2b}\right) + \frac{\pi}{2}\right] \\
&= \frac{1}{2} - \frac{1}{\pi}\tan^{-1}\left|\frac{a_1-a_2}{b}\right|.
\end{aligned}$$

Since $a_1 < a_2$, we have $a_1 - a_2 < 0$, and so $a_1 - a_2 = -|a_1 - a_2|$ and

$$\tan^{-1}\left(\frac{a_1-a_2}{b}\right) = \tan^{-1}\left(-\left|\frac{a_1-a_2}{b}\right|\right) = -\tan^{-1}\left|\frac{a_1-a_2}{b}\right|.$$

7. Find the discriminant function for two class classification problem where the feature vectors are binary and independent given the class? (Assume 0-1 loss)

Solution:

Let the feature vector be $x = (x_1, \ldots, x_d)$ and the two classes be w_1 and w_2.

Further, since the features are binary, let $P(x_i = 1|w_1) = p_i$ and $P(x_i = 1|w_2) = q_i$.

Due to conditional independence, we have

$$P(x|w_1) = \prod_{i=1}^{d} P(x_i|w_1) = \prod_{i=1}^{d} p_i^{x_i}(1-p_i)^{1-x_i}.$$

Similarly, $P(x|w_2) = \prod_{i=1}^{d} q_i^{x_i}(1-q_i)^{1-x_i}$.

The discriminant function for each class can be the sum of log likelihood and log prior.

$$g_1(x) = \sum_{i=1}^{d} [x_i \log p_i + (1 - x_i) \log(1 - p_i)] + \log P(w_1)$$
$$g_2(x) = \sum_{i=1}^{d} [x_i \log q_i + (1 - x_i) \log(1 - q_i)] + \log P(w_2)$$

Since there are only two classes, we can write a single discriminant function as the different between the above two, as

$$\begin{aligned} g(x) &= g_1(x) - g_2(x) \\ &= \sum_{i=1}^{d} \left[x_i \log \frac{p_i}{q_i} + (1 - x_i) \log \frac{1 - p_i}{1 - q_i} \right] + \log \frac{P(w_1)}{P(w_2)} \\ &= \sum_{i=1}^{d} \left(\log \frac{p_i(1 - q_i)}{q_i(1 - p_i)} \right) x_i + \sum_{i=1}^{d} \log \frac{1 - p_i}{1 - q_i} + \log \frac{P(w_1)}{P(w_2)}, \end{aligned}$$

which is a linear function of x.

8. Let $\omega_{max}(x)$ be the state of nature for which $P(\omega_{max}|x) \geq P(\omega_i|x)$ for all i, $i = 1, ..., c$. (Duda Hart Prob 12)
   (a) Show that $P(\omega_{max}|x) \geq \frac{1}{c}$?
   (b) Show that for the minimum-error-rate decision rule the average probability of error is given by

$$P(error) = 1 - \int P(\omega_{max}|x)p(x)dx.$$

Solution:
(a)

$$\begin{aligned} 1 &= \sum_i P(w_i|x) \\ &\leq \sum_i P(w_{max}|x) \\ &= P(w_{max}|x) \sum_i 1 \\ &= P(w_{max}|x)c, \end{aligned}$$

and so $P(w_{max}|c) \geq \frac{1}{c}$.

(b)

$$P(error) = \int_x P(error, x)dx$$
$$= \int_x P(error|x)P(x)dx.$$

Using a minimum error rate decision rule, the class with the highest posterior probability is chosen, so the probability of error is $1-P(w_{max}|x)$. Therefore,

$$P(error) = \int_x (1 - P(w_{max}|x))\, P(x)dx$$
$$= \int_x P(x)dx - \int_x P(w_{max}|x)p(x)dx$$
$$= 1 - \int_x P(w_{max}|x)p(x)dx.$$

9. Consider a simple linear regression model in which y is the sum of a deterministic linear function of x, plus random noise η.

$$y = wx + \eta$$

where x is the real-valued input; y is the real-valued output; and w is a single real-valued parameter to be learned. Here η is a real-valued random variable that represents noise, and that follows a Gaussian distribution with mean 0 and standard deviation σ; that is, $\eta \sim N(0, \sigma)$.
Find the MAP estimate for parameter w assuming a gaussian prior with variance τ
$http : //www.cs.cmu.edu/\ tom/10701\_sp11/midterm.pdf$
Solution:

$$\begin{aligned} p(w|\mathcal{Y}, \mathcal{X}) &\propto p(\mathcal{Y}|\mathcal{X}, w)p(w|\mathcal{X}) \\ &\propto \exp\left\{-\frac{\sum_{i=1}^{n}(y^i - wx^i)^2}{2\sigma^2}\right\} \exp\left\{-\frac{w^2}{2\tau^2}\right\} \end{aligned}$$

$$\begin{aligned} w^* &= \text{argmax}_w \ln p(w|\mathcal{Y}, \mathcal{X}) \\ &= \text{argmax}_w - \frac{\sum_{i=1}^{n}(y^i - wx^i)^2}{2\sigma^2} - \frac{w^2}{2\tau^2} \\ &= \text{argmin}_w \frac{\sum_{i=1}^{n}(y^i - wx^i)^2}{2\sigma^2} + \frac{w^2}{2\tau^2} \\ &= \text{argmin}_w \frac{1}{2} \sum_{i=1}^{n}(y^i - wx^i)^2 + \frac{\sigma^2}{2\tau^2} w^2 \end{aligned}$$