Weekly Overview Slides of Statistical Machine Learning CSE 575, Fall 2023

Moses A. Boudourides[1]

SPA and SCAI
Arizona State University

[1] Moses.Boudourides@asu.edu

Week 6

Exercises on Linear Basis Function Models, Discriminant Functions, and Probabilistic Generative Models

February 16, 2023

Exercise 1

The logistic sigmoid function in terms of tanh

Show that the logistic sigmoid function $\sigma(\alpha) = 1/\big(1 + \exp(-\alpha)\big)$ satisfies the equation $2\sigma(2\alpha) - 1 = \tanh$ and find the corresponding form to the general linear combination of logistic sigmoid function $y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + \sum_{j=1}^{M} w_j \sigma\big(\frac{x-\mu_j}{s}\big)$.

Solution

$$2\sigma(2\alpha) - 1 = \frac{2}{1 + e^{-2\alpha}} - 1 = \frac{2}{1 + e^{-2\alpha}} - \frac{1 + e^{-2\alpha}}{1 + e^{-2\alpha}}$$

$$= \frac{1 - e^{-2\alpha}}{1 + e^{-2\alpha}} = \frac{e^{\alpha} - e^{-\alpha}}{e^{\alpha} + e^{-\alpha}} = \tanh(\alpha).$$

Thus, taking $\alpha_j = (x - \mu_j)/(2s)$, we get

$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + \sum_{j=1}^{M} w_j \sigma\big(2\alpha_j\big) = w_0 + \sum_{j=1}^{M} \frac{w_j}{2}\big(2\sigma(2\alpha_j) - 1 + 1\big)$$

$$= u_0 + \sum_{j=1}^{M} u_j \tanh(\alpha_j),$$

where $u_j = w_j/2$, for $j = 1, \ldots, M$, and $u_0 = w_0 + \sum_{j=1}^{M} w_j/2$.

Exercise 2

Orthogonal Projections in Linear Algebra

If the columns of a matrix $\mathbf{\Phi}$ are independent, show that the matrix $\mathbf{\Phi}\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T$ takes any vector $\boldsymbol{v}$ and projects it onto the space spanned by the columns of $\mathbf{\Phi}$.

Solution

Since the columns of $\mathbf{\Phi}$ are independent, one can show by contradiction that $\mathbf{\Phi}^T\mathbf{\Phi}$ is invertible. Because, if $\mathbf{\Phi}^T\mathbf{\Phi}$ was not invertible, then there would exist a vector $\boldsymbol{x} \neq \mathbf{0}$ such that $\mathbf{\Phi}^T\mathbf{\Phi}\boldsymbol{x} = \mathbf{0}$. So, $\boldsymbol{x}^T(\mathbf{\Phi}^T\mathbf{\Phi}\boldsymbol{x}) = 0$, i.e., $(\mathbf{\Phi}\boldsymbol{x})^T(\mathbf{\Phi}\boldsymbol{x}) = 0$, which would mean that $||\mathbf{\Phi}\boldsymbol{x}|| = 0$, i.e., $\mathbf{\Phi}\boldsymbol{x} = \mathbf{0}$, and, thus, the columns of $\mathbf{\Phi}$ could not be independent, which is a contradiction. Moreover, the invertibility of $\mathbf{\Phi}^T\mathbf{\Phi}$ implies that $(\mathbf{\Phi}^T\mathbf{\Phi})^{-1} = \mathbf{\Phi}^{-1}\mathbf{\Phi}^{-T}$. Next, let us consider the matrix

$$\mathbf{\Psi} = \mathbf{\Phi}\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T.$$

Then it is easy to see that both the following two conditions hold

$$\mathbf{\Psi}^T = \mathbf{\Psi}$$

$$\mathbf{\Psi}^2 = \mathbf{\Psi}.$$

Therefore, $\mathbf{\Psi}$ is a projection matrix on the range of $\mathbf{\Phi}$, which is what we want.

Exercise 3

A Linear Model

Consider a linear model of the form

$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + \sum_{i=1}^{D} w_i x_i, \tag{1}$$

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

$$E_D(\boldsymbol{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \boldsymbol{w}) - t_n\}^2 \tag{2}$$

Now suppose that Gaussian noise ϵ_i with zero mean and variance σ^2 is added independently to each of the input variables x_i. By making use of $\mathbb{E}[\epsilon_i] = 0$ and $\mathbb{E}[\epsilon_i \epsilon_j] = \delta_{ij}\sigma^2$, 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.

Solution of Exercise 3

Setting

$$\tilde{y}_n = w_0 + \sum_{i=1}^{D} w_i(x_{ni} + \epsilon_{ni})$$

$$= y_n + \sum_{i=1}^{D} w_i \epsilon_{ni},$$

where $y_n = y(x_n, \boldsymbol{w})$ and $\epsilon_{ni} \sim \mathcal{N}(0, \sigma^2)$, after using (1). Thus, as in (2), we define

$$\tilde{E} = \frac{1}{2}\sum_{n=1}^{N}\{\tilde{y}_n - t_n\}^2 = \frac{1}{2}\sum_{n=1}^{N}\{\tilde{y}_n^2 - 2\tilde{y}_n t_n + t_n^2\}$$

$$= \frac{1}{2}\sum_{n=1}^{N}\left\{y_n^2 + 2y_n\sum_{i=1}^{D} w_i\epsilon_i + \left(\sum_{i=1}^{D} w_i\epsilon_i\right)^2 - 2t_n y_n - 2t_n\sum_{i=1}^{D} w_i\epsilon_{ni} + t_n^2\right\}.$$

Taking the expectation of $\tilde{E}$ under the distribution of ϵ_{ni}, the second and the fifth term disappear, since $\mathbb{E}[\epsilon_{ni}] = 0$, while the third term becomes

$$\mathbb{E}\left[\left(\sum_{n=1}^{D} w_i\epsilon_{ni}\right)^2\right] = \sum_{i=1}^{D} w_i^2\sigma^2,$$

since the ϵ_{ni} are all independent with variance σ^2. From this and (2), we obtain

$$\mathbb{E}\left[\tilde{E}\right] = E_D + \frac{1}{2}\sum_{i=1}^{D} w_i^2\sigma^2.$$

Exercise 4

Lagrange Multipliers

Using the technique of Lagrange multipliers, show that the minimization of the regularized error function (3.29) is equivalent to minimizing the uregularized sum–of–squares error (3.12) subject to the constraint

$$\sum_{j=1}^{M} |w_j|^q \leq \eta.$$

Solution

Apparently, the above constraint can be written as $\frac{1}{2}\left(\sum_{j=1}^{M} |w_j|^q - \eta\right) \leq 0$, which combined with (3.12) yields the Lagrange function

$$L(\boldsymbol{w}, \lambda) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \boldsymbol{w}^T\phi(\boldsymbol{x}_n)\}^2 + \frac{\lambda}{2}\left(\sum_{j=1}^{M} |w_j|^q - \eta\right),$$

which is identical to (3.29) in the dependence on $\boldsymbol{w}$. Now, let us choose a specific value of $\lambda > 0$ to minimize (3.29). Denoting the resulting value of $\boldsymbol{w}$ as $\boldsymbol{w}^*(\lambda)$, and using the KKT condition $\lambda g(\boldsymbol{w}) = 0$, we get the following value of η

$$\eta = \sum_{j=1}^{M} |w_j^*(\lambda)|^q.$$

Exercise 5

A Linear Basis Function Regression Gaussian Model

Consider a linear basis function regression model for a multivariate target variable $\boldsymbol{t}$ having a Gaussian distribution of the form

$$p(\boldsymbol{t}|\boldsymbol{W}, \Sigma) = \mathcal{N}(\boldsymbol{t}|\boldsymbol{y}(\boldsymbol{x}, \boldsymbol{W}), \Sigma),$$

where

$$\boldsymbol{y}(\boldsymbol{x}, \boldsymbol{W}) = \boldsymbol{W}^T \phi(\boldsymbol{x}),$$

together with a training data set comprising input basis vectors $\phi(\boldsymbol{x}_n)$ and corresponding target vectors $\boldsymbol{t}_n$, with $n = 1, \ldots, N$. Show that the maximum likelihood solution $\boldsymbol{W}_{\mathrm{ML}}$ for the parameter matrix $\boldsymbol{W}$ has the property that each column is given by an expression of the form (3.15), which was the solution for an isotropic noise distribution. Note that this is independent of the covariance matrix Σ. Show that the maximum likelihood solution for Σ is given by

$$\Sigma = \frac{1}{N} \sum_{n=1}^{N} \left(\boldsymbol{t}_n - \boldsymbol{W}_{\mathrm{ML}}^T \phi(\boldsymbol{x}_n)\right) \left(\boldsymbol{t}_n - \boldsymbol{W}_{\mathrm{ML}}^T \phi(\boldsymbol{x}_n)\right)^T.$$

Solution of Exercise 5

The corresponding log–likelihood function is

$$\log L(\boldsymbol{W}, \boldsymbol{\Sigma}) = -\frac{N}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{n=1}^{N} \Big(\boldsymbol{t}_n - \boldsymbol{W}^T \phi(\boldsymbol{x}_n)\Big)^T \boldsymbol{\Sigma}^{-1} \Big(\boldsymbol{t}_n - \boldsymbol{W}^T \phi(\boldsymbol{x}_n)\Big).$$

Setting the $\boldsymbol{W}$ derivative equal to 0 yields

$$0 = -\sum_{n=1}^{N} \boldsymbol{\Sigma}^{-1} \Big(\boldsymbol{t}_n - \boldsymbol{W}^T \phi(\boldsymbol{x}_n)\Big) \phi(\boldsymbol{x}_n)^T.$$

Multiplying through by $\boldsymbol{\Sigma}$ and introducing the design matrix $\boldsymbol{\Phi}$ and the target data matrix $\boldsymbol{T}$ we have

$$\boldsymbol{\Phi}^T \boldsymbol{\Phi} \boldsymbol{W} = \boldsymbol{\Phi}^T \boldsymbol{T},$$

which solved for $\boldsymbol{\Phi}$ gives

$$\boldsymbol{W}_{\mathrm{ML}} = \Big(\boldsymbol{\Phi}^T \boldsymbol{\Phi}\Big)^{-1} \boldsymbol{\Phi}^T \boldsymbol{T},$$

where the operator $\Big(\boldsymbol{\Phi}^T \boldsymbol{\Phi}\Big)^{-1} \boldsymbol{\Phi}^T = \boldsymbol{\Phi}^\dagger$ is the *Moore–Penrose pseudo-inverse* of the matrix $\boldsymbol{\Phi}$. Finally, the maximum likelihood solution for $\boldsymbol{\Sigma}$ is found as in section 3.1 (by appealing to the standard result from Chapter 2).

Exercise 6

Convex Hulls and Linear Separability

Given a set of data points $\{\boldsymbol{x}_n\}$, we can define the *convex hull* to be the set of all points $\boldsymbol{x}$ given by

$$\boldsymbol{x} = \sum_n \alpha_n \boldsymbol{x}_n,$$

where $\alpha_n \geq 0$ and $\sum_n \alpha_n = 1$. Consider a second set of points $\{\boldsymbol{y}_n\}$ together with their corresponding convex hull. By definition, the two sets of points will be *linearly separable* if there exists a vector $\hat{\boldsymbol{w}}$ and a scalar w_0 such that $\hat{\boldsymbol{w}}^T \boldsymbol{x}_n + w_0 > 0$, for all $\boldsymbol{x}_n$, and $\hat{\boldsymbol{w}}^T \boldsymbol{y}_n + w_0 < 0$, for all $\boldsymbol{y}_n$. Show that if their convex hulls intersect, the two sets of points cannot be linearly separable, and conversely that if they are linearly separable, their convex hulls do not intersect.

Solution of Exercise 6

If the convex hulls of $\{\boldsymbol{x}_n\}$ and $\{\boldsymbol{y}_n\}$ intersect, there exist a point $\boldsymbol{z}$ such that $\boldsymbol{z} = \sum_n \alpha_n \boldsymbol{x}_n = \sum_n \beta_n \boldsymbol{y}_n$. Hence, using that the coefficients of a convex hull sum up to 1, we would have

$$\hat{\boldsymbol{w}}^T \boldsymbol{z} + w_0 = \hat{\boldsymbol{w}}^T \left(\sum_n \alpha_n \boldsymbol{x}_n \right) + w_0 = \left(\sum_n \alpha_n \hat{\boldsymbol{w}}^T \boldsymbol{x}_n \right) + \left(\sum_n \alpha_n \right) w_0$$

$$= \sum_n \alpha_n (\hat{\boldsymbol{w}}^T \boldsymbol{x}_n + w_0).$$

Now, if $\{\boldsymbol{x}_n\}$ and $\{\boldsymbol{y}_n\}$ were linearly separable, we would have $\hat{\boldsymbol{w}}^T \boldsymbol{x}_n + w_0 > 0$, for all $\boldsymbol{x}_n$, and $\hat{\boldsymbol{w}}^T \boldsymbol{y}_n + w_0 < 0$, for all $\boldsymbol{y}_n$. Clearly, the first of the latter conditions would imply (using the previous expression) that $\hat{\boldsymbol{w}}^T \boldsymbol{z} + w_0 > 0$. Similarly, from the corresponding properties of $\{\boldsymbol{y}_n\}$, we would get that $\hat{\boldsymbol{w}}^T \boldsymbol{z} + w_0 < 0$, which is a contradiction and, therefore, the required result follows.

Exercise 7

Least Squares for Classification

Consider the minimization of a sum-of-squares error function $E_D(\tilde{\boldsymbol{W}}) = \frac{1}{2}\mathrm{Tr}\{(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{W}} - \boldsymbol{T})^T(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{W}} - \boldsymbol{T})\}$ $(*)$, and suppose that all of the target vectors in the training set satisfy a linear constraint

$$\boldsymbol{a}^T\boldsymbol{t}_n + b = 0,$$

where $\boldsymbol{t}_n$ corresponds to the n–th row of the matrix $\boldsymbol{T}$ in $(*)$. Show that, as a consequence of this constraint, the elements of the model prediction $\boldsymbol{y}(\boldsymbol{x})$ given by the least-squares solution $\boldsymbol{y}(\boldsymbol{x}) = \tilde{\boldsymbol{W}}^T\tilde{\boldsymbol{x}} = \boldsymbol{T}^T\left(\tilde{\boldsymbol{X}}^\dagger\right)^T\tilde{\boldsymbol{x}}$ (where $\tilde{\boldsymbol{X}}^\dagger$ is the pseudo–inverse of the matrix $\tilde{\boldsymbol{X}}$) also satisfy this constraint, so that

$$\boldsymbol{a}^T\boldsymbol{y}(\boldsymbol{x}) + b = 0.$$

To do so, assume that one of the basis functions $\phi_0(\boldsymbol{x}) = 1$ so that the corresponding parameter w_0 plays the role of a bias.

Solution of Exercise 7

Remember that in the N–dimenional classification case, since $\tilde{\boldsymbol{W}} = [\boldsymbol{w}_0, \boldsymbol{W}^T]^T$ and $\tilde{\boldsymbol{X}} = [\boldsymbol{1}, \boldsymbol{X}^T]^T$, we get the following explicit dependence of $E_D(\tilde{\boldsymbol{W}})$ on $\boldsymbol{w}_0$:

$$E_D(\tilde{\boldsymbol{W}}) = \frac{1}{2}\mathrm{Tr}\Big\{(\boldsymbol{X}\boldsymbol{W} + \boldsymbol{1}\boldsymbol{w}_0^T - \boldsymbol{T})^T(\boldsymbol{X}\boldsymbol{W} + \boldsymbol{1}\boldsymbol{w}_0^T - \boldsymbol{T})\Big\}.$$

The above expression is a matrix quadratic w.r.t. $\boldsymbol{w}_0$ and its $\boldsymbol{w}_0$ derivative can be easily found to be:

$$\frac{\partial}{\partial \boldsymbol{w}_0} E_D(\tilde{\boldsymbol{W}}) = 2N\boldsymbol{w}_0 + 2(\boldsymbol{X}\boldsymbol{W} - \boldsymbol{T})^T\boldsymbol{1}.$$

Thus, setting the derivative equal to 0, we get

$$\boldsymbol{w}_0 = -\frac{1}{N}(\boldsymbol{X}\boldsymbol{W} - \boldsymbol{T})^T\boldsymbol{1} = \bar{\boldsymbol{t}} - \boldsymbol{W}^T\bar{\boldsymbol{x}},$$

where

$$\bar{\boldsymbol{t}} = \frac{1}{N}\boldsymbol{T}^T\boldsymbol{1} \text{ and } \bar{\boldsymbol{x}} = \frac{1}{N}\boldsymbol{X}^T\boldsymbol{1}.$$

Substituting the above critical value of $\boldsymbol{w}_0$ in the expression for $E_D(\tilde{\boldsymbol{W}})$, we obtain

$$E_D(\tilde{\boldsymbol{W}}) = \frac{1}{2}\mathrm{Tr}\Big\{(\boldsymbol{X}\boldsymbol{W} + \overline{\boldsymbol{T}} - \overline{\boldsymbol{X}}\boldsymbol{W} - \boldsymbol{T})^T(\boldsymbol{X}\boldsymbol{W} + \overline{\boldsymbol{T}} - \overline{\boldsymbol{X}}\boldsymbol{W} - \boldsymbol{T})\Big\},$$

where we have denoted

$$\overline{\boldsymbol{T}} = \boldsymbol{1}\bar{\boldsymbol{t}}^T \text{ and } \overline{\boldsymbol{X}} = \boldsymbol{1}\bar{\boldsymbol{x}}^T.$$

Solution of Exercise 7 (cont.)

Next, denoting two new matrices

$$\hat{\boldsymbol{X}} = \boldsymbol{X} - \overline{\boldsymbol{X}} \text{ and } \hat{\boldsymbol{T}} = \boldsymbol{T} - \overline{\boldsymbol{T}},$$

setting the $\boldsymbol{W}$ derivative of $E_D(\tilde{\boldsymbol{W}})$ equal to 0 yields

$$\boldsymbol{W} = (\hat{\boldsymbol{X}}^T \hat{\boldsymbol{X}})^{-1} \hat{\boldsymbol{X}}^T \hat{\boldsymbol{T}} = \hat{\boldsymbol{X}}^\dagger \hat{\boldsymbol{T}},$$

where $\hat{\boldsymbol{X}}^\dagger = (\hat{\boldsymbol{X}}^T \hat{\boldsymbol{X}})^{-1} \hat{\boldsymbol{X}}^T$ is the *Moore–Penrose pseudo–inverse* of $\hat{\boldsymbol{X}}$.

Finally, consider the prediction for a new input vector $\boldsymbol{x}^*$

$$\begin{aligned} \boldsymbol{y}(\boldsymbol{x}^*) &= \boldsymbol{W}^T \boldsymbol{x}^* + \boldsymbol{w}_0 = \boldsymbol{W}^T \boldsymbol{x}^* + \overline{\boldsymbol{t}} - \boldsymbol{W}^T \overline{\boldsymbol{x}} \\ &= \overline{\boldsymbol{t}} + \hat{\boldsymbol{T}}^T \left(\hat{\boldsymbol{X}}^\dagger\right)^T (\boldsymbol{x}^* - \overline{\boldsymbol{x}}). \end{aligned}$$

If we know that $\boldsymbol{a}^T \boldsymbol{t}_n + b = 0$ holds, for some $\boldsymbol{a}$ and b, we get

$$\boldsymbol{a}^T \overline{\boldsymbol{t}} = \frac{1}{N} \boldsymbol{a}^T \boldsymbol{T}^T \boldsymbol{1} = -b.$$

Therefore, applying the condition $\boldsymbol{a}^T \boldsymbol{t}_n + b = 0$ to the previously derived expression for $\boldsymbol{x}^*$, we obtain

$$\boldsymbol{a}^T \boldsymbol{y}(\boldsymbol{x}^*) = \boldsymbol{a}^T \overline{\boldsymbol{t}} + \boldsymbol{a}^T \hat{\boldsymbol{T}}^T \left(\hat{\boldsymbol{X}}^\dagger\right)^T (\boldsymbol{x}^* - \overline{\boldsymbol{x}}) = \boldsymbol{a}^T \overline{\boldsymbol{t}} = -b,$$

since $\boldsymbol{a}^T \hat{\boldsymbol{T}}^T = \boldsymbol{a}^T (\boldsymbol{T} - \overline{\boldsymbol{T}})^T = b(\boldsymbol{1} - \boldsymbol{1})^T = \boldsymbol{0}^T$.

Exercise 8

Maximization of the Class Separation Criterion with Lagrange Multiplier

Show that maximization with respect to $\boldsymbol{w}$ of the class separation criterion given in Bishop's PRML book as $m_2 - m_1 = \boldsymbol{w}^T(\boldsymbol{m}_2 - \boldsymbol{m}_1)$, using a Lagrange multiplier to enforce the constraint $\boldsymbol{w}^T\boldsymbol{w} = 1$, leads to the result that $\boldsymbol{w} \propto (\boldsymbol{m}_2 - \boldsymbol{m}_1)$.

Solution

Using a Lagrange multiplier to enforce the constraint $\boldsymbol{w}^T\boldsymbol{w} = 1$ means that we need to maximize the following objective function

$$L(\lambda, \boldsymbol{w}) = \boldsymbol{w}^T(\boldsymbol{m}_2 - \boldsymbol{m}_1) + \lambda(\boldsymbol{w}^T\boldsymbol{w} - 1).$$

Clearly, the $\boldsymbol{w}$ derivative (gradient) of this Lagrange function is

$$\frac{\partial}{\partial \boldsymbol{w}} L(\lambda, \boldsymbol{w}) = \boldsymbol{m}_2 - \boldsymbol{m}_1 + 2\lambda\boldsymbol{w},$$

which set equal to 0 gives

$$\boldsymbol{w} = -\frac{1}{2\lambda}(\boldsymbol{m}_2 - \boldsymbol{m}_1) \propto (\boldsymbol{m}_2 - \boldsymbol{m}_1).$$

Exercise 9

The Alternative Form of the Fisher Criterion

By making use of (4.20), (4.23), and (4.24), show that the Fisher criterion (4.25) can be written in the form (4.26), where the previous parentheses correspond to numbers of euations in Bishop's PRML book.

Solution

First, we expand (4.25) using (4.22), (4.23) and (4.24):

$$J(\boldsymbol{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} = \frac{||\boldsymbol{w}^T(\boldsymbol{m}_2 - \boldsymbol{m}_1)||^2}{\sum_{n\in C_1}(\boldsymbol{w}^T\boldsymbol{x}_n - m_1)^2 + \sum_{n\in C_2}(\boldsymbol{w}^T\boldsymbol{x}_n - m_2)^2}.$$

Next, on the one hand, denoting

$$\boldsymbol{S_B} = (\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^T,$$

the numerator becomes

$$\text{numerator} = \Big(\boldsymbol{w}^T(\boldsymbol{m}_2 - \boldsymbol{m}_1)\Big)\Big(\boldsymbol{w}^T(\boldsymbol{m}_2 - \boldsymbol{m}_1)\Big)^T = \boldsymbol{w}^T\boldsymbol{S_B}\boldsymbol{w}.$$

On the other hand, denoting

$$\boldsymbol{S_w} = \sum_{n\in C_1}(\boldsymbol{x}_n - \boldsymbol{m}_1)(\boldsymbol{x}_n - \boldsymbol{m}_1)^T + \sum_{n\in C_2}(\boldsymbol{x}_n - \boldsymbol{m}_2)(\boldsymbol{x}_n - \boldsymbol{m}_2)^T,$$

the denominator becomes

$$\begin{aligned}\text{denominator} &= \sum_{n\in C_1}\Big(\boldsymbol{w}^T(\boldsymbol{x}_n - \boldsymbol{m}_1)\Big)^2 + \sum_{n\in C_2}\Big(\boldsymbol{w}^T(\boldsymbol{x}_n - \boldsymbol{m}_2)\Big)^2 \\ &= \boldsymbol{w}^T\boldsymbol{S}_{w_1}\boldsymbol{w} + \boldsymbol{w}^T\boldsymbol{S}_{w_2}\boldsymbol{w} = \boldsymbol{w}^T\boldsymbol{S_w}\boldsymbol{w}.\end{aligned}$$

Thus, the wanted result.

Exercise 10

The Logistic Sigmoid Function

Show that the logistic sigmoid function $\sigma(\alpha) = 1/(1 + \exp(-\alpha))$ satisfies the property $\sigma(-\alpha) = 1 - \sigma(\alpha)$ and that its inverse is given by $\sigma^{-1}(y) = \ln\{y/(1-y)\}$.

Solution

By definition of the logistic sigmoid function, we get

$$\begin{aligned}\sigma(\alpha) + \sigma(-\alpha) &= \frac{1}{1+\exp(-\alpha)} + \frac{1}{1+\exp(\alpha)} \\ &= \text{ DO THE ALGEBRA } = 1.\end{aligned}$$

Next, exchanging the dependent and independent variables in $y = \sigma(\alpha)$, we obtain the inverse of the logistic sigmoid function

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

and after rearranging it as

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

if we take the logarithm in both sides, we get

$$y = \ln\{\frac{\alpha}{1-\alpha}\},$$

which is what we wanted (writing $\sigma^{-1}(y)$ instead of y and y instead of α).

Exercise 11

The Two–Class Generative Model with Gaussian Densities

Using (4.57) and (4.58), derive the result (4.65) for the posterior class probability in the two–class generative model with Gaussian densities, and verify the results (4.66) and (4.67) for the parameters $\boldsymbol{w}$ and w_0 (where all parentheses correspond to numbers of euations in Bishop's PRML book).

Solution

According to (4.58) and (4.64), we can write

$$\begin{aligned}
\alpha &= \ln \frac{p(\boldsymbol{x}|C_1)p(C_1)}{p(\boldsymbol{x}|C_2)p(C_2)} = \ln p(\boldsymbol{x}|C_1) - \ln p(\boldsymbol{x}|C_2) + \ln \frac{p(C_1)}{p(C_2)} \\
&= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) + \ln \frac{p(C_1)}{p(C_2)} \\
&= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln \frac{p(C_1)}{p(C_2)} \\
&= \boldsymbol{w}^T\boldsymbol{x} + w_0.
\end{aligned}$$

Notice that, in the last step, we have rearranged the terms, and have set

$$\boldsymbol{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2),$$

$$w_0 = -\frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln \frac{p(C_1)}{p(C_2)}.$$

Thus, since $p(C_1|\boldsymbol{x}) = \sigma(\alpha)$ (as stated in (4.57)), we get $p(C_1|\boldsymbol{x}) = \sigma(\boldsymbol{w}^T\boldsymbol{x} + w_0)$, as required.

Exercise 12

Maximum–Likelihood Solution to a Generative Classification Model

Consider a generative classification model for K classes defined by prior class probabilities $p(C_k) = \pi_k$ and general class–conditional densities $p(\phi|C_k)$ where ϕ is the input feature vector. Suppose we are given a training data set $\{\phi_n, t_n\}$, where $n = 1, \ldots, N$, and t_n is a binary target vector of length K that uses the 1–of–K coding scheme, so that it has components $t_{nj} = I_{jk}$, if pattern n is from class C_k. Assuming that the data points are drawn independently from this model, show that the maximum–likelihood solution for the prior probabilities is given by

$$\pi_k = \frac{N_k}{N},$$

where N_k is the number of data points assigned to class C_k.

Solution of Exercise 12

The likelihood function is given by

$$p(\{\phi_n, \boldsymbol{t}_n\}|\{\pi_k\}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\{p(\phi_n|C_k)\pi_k\}^{t_{nk}}$$

and taking the logarithm we obtain

$$\ln p(\{\phi_n, \boldsymbol{t}_n\}|\{\pi_k\}) = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\{\ln p(\phi_n|C_k) + \ln \pi_k\} \propto \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln \pi_k.$$

Since there is a constraint on π_k, we need to add a Lagrange multiplier to the above expression

$$L = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln \pi_k + \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right).$$

Clearly, the π_k derivative of the Lagrangian function is

$$\frac{\partial}{\partial \pi_k} L = \sum_{n=1}^{N} \frac{t_{nk}}{\pi_k} + \lambda$$

and, setting the derivative equal to 0, we obtain

$$\pi_k = -\frac{\sum_{n=1}^{N} t_{nk}}{\lambda} = -\frac{N_k}{\lambda}.$$

Summing both sides of the last expression with regard to k, we get

$$1 = -\frac{\sum_{n=1}^{N} N_k}{\lambda} = -\frac{N_k}{\lambda},$$

i.e., $\lambda = -N$, which implies that $\pi_k = N_k/N$.