16.5 Exercises Lecture 5

16.5.1 Exercise 1: A generative 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 $n = 1,...,N$ 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

$$\pi_k = \frac{N_k}{N} \tag{16.91}$$

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

16.5.2 Solution

We begin by writing down the likelihood function.

$$p(\{\boldsymbol{\phi_n}, t_n\}|\pi_1, \pi_2, ..., \pi_K) = \prod_{n=1}^{N}\prod_{k=1}^{K}[p(\boldsymbol{\phi_n}|C_k)p(C_k)]^{t_{nk}} = \prod_{n=1}^{N}\prod_{k=1}^{K}[\pi_k p(\boldsymbol{\phi_n}|C_k)]^{t_{nk}} \tag{16.92}$$

Hence we can obtain the expression for the logarithm likelihood:

$$ln\ p = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}[ln\pi_k + ln\ p(\boldsymbol{\phi_n}|C_k)] \propto \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} ln\pi_k \tag{16.93}$$

Since there is a constraint on π_k, so we need to add a Lagrange Multiplier to the expression, which becomes:

$$L = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} ln\pi_k + \lambda(\sum_{k=1}^{K}\pi_k - 1) \tag{16.94}$$

We calculate the derivative of the expression above with regard to π_k:

$$\frac{\partial L}{\partial \pi_k} = \sum_{n=1}^{N}\frac{t_{nk}}{\pi_k} + \lambda \tag{16.95}$$

And if we set the derivative equal to 0, we can obtain:

$$\pi_k = -(\sum_{n=1}^{N} t_{nk}/\lambda = -\frac{N_k}{\lambda} \tag{16.96}$$

And if we preform summation on both sides with regard to k, we can see that:

$$1 = -(\sum_{k=1}^{K} N_k)/\lambda = -\frac{N_k}{\lambda} \tag{16.97}$$

Which gives $\lambda = -N$, and substitute it into 16.96, we can obtain the result.

16.5.3 Exercise 2: An example of Naive Bayes model

Consider a classification problem with K classes for which the feature vector ϕ has M components each of which can take L discrete states. Let the values of the components be represented by a 1-of-L binary coding scheme. Further suppose that, conditioned on the class C_k, the M components of ϕ are independent, so that the class-conditional density factorizes with respect to the feature vector components. Show that the quantities a_k given by :

$$a_k = ln\ p(x|C_k)p(C_k) \tag{16.98}$$

which appear in the argument to the softmax function describing the posterior class probabilities, are linear functions of the components of ϕ. Note that this represents an example of the naive Bayes model

16.5.4 Solution

Based on definition, we can write down

$$p(\boldsymbol{\phi}, C_k) = \prod_{m=1}^{M} \prod_{l=1}^{L} \mu_{kml}^{\phi_{ml}} \tag{16.99}$$

Note that here only one of the value among $\phi_{m1}, \phi_{m2}, ...\phi_{mL}$ is 1, and the others are all 0 because we have used a $1-of-L$ binary coding scheme, and also we have taken advantage of the assumption that the M components of ϕ are independent conditioned on the class C_k. We substitute the expression above into

$$a_k = lnp(\boldsymbol{x}|C_k)p(C_k) \tag{16.100}$$

which gives:

$$a_k = \sum_{m=1}^{M} \sum_{l=1}^{L} \phi_{ml} \cdot ln\ \mu_k ml + lnp(C_k) \tag{16.101}$$

Hence it is obvious that a_k is a linear function of the components of ϕ

16.5.5 Exercise 3: A follow-up of exercise 1

Consider the classification model of exercise 1 and now suppose that the class-conditional densities are given by Gaussian distributions with a shared covariance matrix, so that

$$p(\phi, C_k) = N(\phi|\mu_k, \Sigma) \tag{16.102}$$

Show that the maximum likelihood solution for the mean of the Gaussian distribution for class C_k is given by

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} t_{nk} \phi_n \tag{16.103}$$

which represents the mean of those feature vectors assigned to class C_k. Similarly, show that the maximum likelihood solution for the shared covariance matrix is given by

$$\Sigma = \sum_{k=1}^{K} \frac{N_k}{N} \boldsymbol{S}_k \tag{16.104}$$

where

$$S_k = \frac{1}{N_k} \sum_{n=1}^{N} t_{nk}(\phi_n - \mu_k)(\phi_n - \mu_k)^T \tag{16.105}$$

Thus Σ is given by a weighted average of the covariances of the data associated with each class, in which the weighting coefficients are given by the prior probabilities of the classes.

16.5.6 Exercise 4: Softmax and Sigmoid

Show that the softmax function is equivalent to a sigmoid in the 2-class case

16.5.7 Solution

$$\frac{exp(\boldsymbol{w_1^T x})}{exp(\boldsymbol{w_1^T x}) + exp(\boldsymbol{w_0^T x})} = \frac{1}{1 + exp(\boldsymbol{w_0^T x})/exp(\boldsymbol{w_1^T x})} = \frac{1}{1 + exp(\boldsymbol{w_0^T x} - \boldsymbol{w_1^T x})} =$$
$$= \frac{1}{1 + exp(-(\boldsymbol{w_1^T} - \boldsymbol{w_0^T})^T \boldsymbol{x})} = \sigma(\boldsymbol{w^T x}) \tag{16.106}$$

16.5.8 Exercise 5: A bit of algebra

Show that the derivative of the sigmoid function $\sigma(a) = (1 + e^{-a})^{-1}$ can be written as:

$$\frac{\partial \sigma(a)}{\partial a} = \sigma(a)(1 - \sigma(a)) \tag{16.107}$$

16.5.9 solution

$$\frac{\partial \sigma(a)}{\partial a} = -\frac{1}{(1 + e^{-a})^2} \cdot e^{-a} \cdot (-1) = \frac{1}{1 + e^{-a}} \frac{e^{-a}}{1 + e^{-a}} = \sigma(a) \frac{1 + e^{-a} - 1}{1 + e^{-a}} = \sigma(a)(1 - \sigma(a))$$

(16.108)