Probabilistic Machine Learning: Lecture 2

Isabel Valera

Exercise 1: Sigmoid: the beginning

Show that the logistic sigmoid function given in 1 satisfies the following properties:

i) $\sigma(-a) = 1 - \sigma(a)$

ii) $\sigma^{-1}(y) = \ln \frac{y}{1-y}$

$$\sigma(a) = \frac{1}{1 + \exp -a} \tag{1}$$

Exercise 2: Sigmoid: the posterior

Use equation 2 where a is given by 4.

i) Derive the result of equation 5 for the posterior class probability in the two-class generative model with Gaussian densities.

ii) Verify the results 6 and 7 for the parameters $\boldsymbol{w}$ and w_0.

$$p(\mathcal{C}_1 \,|\, \boldsymbol{x}) = \frac{p(\boldsymbol{x} \,|\, \mathcal{C}_1)p(\mathcal{C}_1)}{p(\boldsymbol{x} \,|\, \mathcal{C}_1)p(\mathcal{C}_1) + p(\boldsymbol{x} \,|\, \mathcal{C}_2)p(\mathcal{C}_2)} \tag{2}$$

$$= \frac{1}{1 + \exp -a} = \sigma(a) \tag{3}$$

$$a = \ln \frac{p(\boldsymbol{x} \,|\, \mathcal{C}_1)p(\mathcal{C}_1)}{p(\boldsymbol{x} \,|\, \mathcal{C}_2)p(\mathcal{C}_2)} \tag{4}$$

$$p(\mathcal{C}_1 \,|\, \boldsymbol{x}) = \sigma(\boldsymbol{w}^\top \boldsymbol{x} + w_0) \tag{5}$$

$$\boldsymbol{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{6}$$

$$w_0 = -\frac{1}{2}\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln \frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)} \tag{7}$$

Exercise 3: One-of-K (1)

Consider a generative classification model for K classes defined by prior class probabilities $p(\mathcal{C}) = \pi_k$ and general class-conditional densities $p(\boldsymbol{\varphi}|\mathcal{C}_k)$ where $\boldsymbol{\varphi}$ is the input feature vector. Suppose we are given a training data set $\{\boldsymbol{\varphi}_n, \boldsymbol{y}_n\}$ where $n = 1, ..., N$ and $\boldsymbol{y}_n$ is a binary target vector of length K that uses the 1-of-K coding scheme, so that it has components $y_{nj} = \boldsymbol{I}_{jk}$ if pattern n is from class $\mathcal{C}_k$. Assuming that the data points are drawn independently from this model, show that the maximum-likelihood solution for the prior probabilities is $\pi_k = \frac{N_k}{N}$ where N_k is the number of data points assigned to class $\mathcal{C}_k$.

Exercise 4: One-of-K (2)

Consider the classification model of exercise 3 and now suppose that the class-conditional densities are given by Gaussian distributions with a shared covariance matrix so that $p(\boldsymbol{\varphi}|\mathcal{C}_k) = \text{Gauss}(\boldsymbol{\varphi}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$.

i) Show that the maximum likelihood solution for the mean of the Gaussian distribution for class $\mathcal{C}_k$, which represents the mean of those feature vectors assigned to class $\mathcal{C}_k$ is given by equation 8

ii) Show that the maximum likelihood solution for the shared covariance matrix $\boldsymbol{\Sigma}$ is given by equation 9. Thus $\boldsymbol{\Sigma}$ is given by a weighted average of the covariances of the data associated with each class given in equation 10, in which the weighting coefficients are given by the prior probabilities of the classes.

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} y_{nk} \boldsymbol{\varphi}_n \tag{8}$$

$$\boldsymbol{\Sigma} = \sum_{k=1}^{K} \frac{N_k}{N} \boldsymbol{S}_k \tag{9}$$

$$\boldsymbol{S}_k = \frac{1}{N_k} \sum_{n=1}^{N} y_{nk} (\boldsymbol{\varphi}_n - \boldsymbol{\mu}_k)(\boldsymbol{\varphi}_n - \boldsymbol{\mu}_k)^\top \tag{10}$$

Exercise 5: Sigmoid: the derivative

Verify the relation given in 11 for the derivative of the logistic sigmoid function defined by equation 1.

$$\frac{d\sigma}{da} = \sigma(1 - \sigma) \tag{11}$$

Exercise 6: Sigmoid: the error

By making use of the result 11 for the derivative of the logistic sigmoid, show that the derivative of the error function given in 12 for the logistic regression model is given by equation 13.

$$L(\boldsymbol{\Phi}, \boldsymbol{y}, \boldsymbol{w}) = -\ln p(\boldsymbol{y} \mid \boldsymbol{w}) = \sum_{n=1}^{N} y_n \ln \sigma(\boldsymbol{w}^\top \boldsymbol{\varphi}_n) + (1 - y_n) \ln(1 - \sigma(\boldsymbol{w}^\top \boldsymbol{\varphi}_n)) \tag{12}$$

$$\nabla_{\boldsymbol{w}} L(\boldsymbol{\Phi}, \boldsymbol{y}, \boldsymbol{w}) = \sum_{n=1}^{N} (\sigma(\boldsymbol{w}^\top \boldsymbol{\varphi}_n) - y_n)\boldsymbol{\varphi}_n \tag{13}$$

Exercise 7: Linearly separable

Show that for a linearly separable data set, the maximum likelihood solution for the logistic regression model is obtained by finding a vector $\boldsymbol{w}$ whose decision boundary $\boldsymbol{w}^\top \varphi(\boldsymbol{x}) = 0$ separates the classes and then taking the magnitude of $\boldsymbol{w}$ to infinity.

Exercise 8: Linear Discriminant Analysis

Suppose we have features $\boldsymbol{x} \in \mathbb{R}^p$, a two-class response, with class sizes N_1, N_2, and the target coded as $-\frac{N}{N_1}, \frac{N}{N_2}$

i) Show that the LDA rule classifies to class 2 if the equation below holds and to class 1 otherwise.

$$\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) > \frac{1}{2}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) - \ln \frac{N_2}{N_1} \tag{14}$$

ii) Consider minimization of the least squares criterion $\sum_{i=1}^{N}(y_i - w_0 - \boldsymbol{x}_i \boldsymbol{w})^2$. Show that the solution w satisfies

$$\begin{aligned} ((N-2)\boldsymbol{\Sigma} + N\boldsymbol{\Sigma}_B)\,\boldsymbol{w} &= N(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \\ \text{where } \boldsymbol{\Sigma}_B &= \frac{N_1 N_2}{N}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top \end{aligned}$$

iii) Hence show that $\boldsymbol{\Sigma}_B \boldsymbol{w}$ is in the direction $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$ and thus $\boldsymbol{w} \propto \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$. Therefore the least-squares regression coefficient is identical to the LDA coefficient, up to a scalar multiple.

iv) Show that this result holds for any (distinct) coding of the two classes.

v) Find the solution w_0 (up to the same scalar multiple as before), and hence the predicted value $f(\boldsymbol{x}) = w_0 + \boldsymbol{x}^\top \boldsymbol{w}$. Consider the following rule: classify to class 2 if $f(\boldsymbol{x}) > 0$ and class 1 otherwise. Show this is the not the same as the LDA rule unless the classes have equal numbers of observations.