Homework 3

Deadline: November 21, 13:15

Exercise 1

Suppose we collect data from a group of students in a Machine learning class with variables x_1 = hours studied, x_2 = grade point average, and $y = a$ binary output if that student received grade 5 ($y = 1$) or not ($y = 0$). We learn a logistic regression model

$$p(y = 1 \mid \boldsymbol{x}) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}$$

with parameters $\hat{\beta}_0 = -6, \hat{\beta}_1 = 0.05, \hat{\beta}_2 = 1$.

(i) Estimate the probability according to the logistic regression model that a student who studies for 40 h and has the grade point average of 3.5 gets a 5 in the Machine learning class.

(ii) According to the logistic regression model, how many hours would the student in part (i) need to study to have 50% chance of getting a 5 in the class?

Exercise 2

(i) Let $\sigma(a) = \frac{1}{1+e^{-a}}$ be the sigmoid function. Show that

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

(ii) Using the previous result and the chain rule of calculus, derive an expression for the gradient of the log likelihood for logistic regression.

(iii) The Hessian can be written as $\boldsymbol{H} = \boldsymbol{X}^T\boldsymbol{S}\boldsymbol{X}$, where $\boldsymbol{S} = \mathrm{diag}(\mu_1(1-\mu_1), \cdots \mu_n(1-\mu_n))$. Show that $\boldsymbol{H}$ is positive definite. (You may assume that $0 < \mu_i < 1$, so the elements of $\boldsymbol{S}$ will be strictly positive, and that $\boldsymbol{X}$ is full rank.)

Exercise 3

We want to create a generative binary classification model for classifying non-negative one-dimensional data. This means, that the labels are binary ($y \in \{0, 1\}$) and the samples are $x \in [0, \infty)$. We assume uniform class probabilities

$$p(y = 0) = p(y = 1) = \frac{1}{2}$$

As our samples x are non-negative, we use exponential distributions (and not Gaussians) as class conditionals:

$$p(x \mid y = 0) = Expo(x \mid \lambda_0) \qquad \text{and} \qquad p(x \mid y = 1) = Expo(x \mid \lambda_1)$$

where $\lambda_0 \neq \lambda_1$. Assume, that the parameters λ_0 and λ_1 are known and fixed.

1. What is the name of the posterior distribution $p(y \mid x)$? You only need to provide the name of the distribution (e.g., "normal", "gamma", etc.), not estimate its parameters.

2. What values of xare classified as class 1? (As usual, we assume that the classification decision is $\hat{y} = argmax_k p(y = k \mid x)$)

Exercise 4

Consider a generative classification model for C classes defined by class probabilities $p(y = c) = \boldsymbol{\pi}_c$ and general class-conditional densities $p(\boldsymbol{x} \mid y = c, \boldsymbol{\theta_c})$, where $\boldsymbol{x} \in \mathbb{R}^D$ is the input feature vector and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_c\}_{c=1}^{C}$. are further model parameters. Suppose we are given a training set $\mathcal{D} = \{(\boldsymbol{x}^n, y^n)\}_{n=1}^{N}$, where $y^{(n)}$ is a binary target vector of length C that uses the 1-of-C (one-hot) encoding scheme, so that it has components $y_c^n = \delta_{ck}$ if pattern n is from class $y = k$. Assuming that the data points are i.i.d.,show that the maximum-likelihood solution for the class probabilities $\boldsymbol{\pi}$ is given by

$$\pi_c = \frac{N_c}{N}$$

where N_c is the number of data points assigned to class c.

Exercise 5

Using the same classification model as in the previous question, now suppose that the class-conditional densities are given by Gaussian distributions with a shared covariance matrix, so that

$$p(\boldsymbol{x} \mid y = c, \boldsymbol{\theta}) = p(\boldsymbol{x} \mid \boldsymbol{\theta_c}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu_c}, \boldsymbol{\Sigma})$$

Show that the maximum likelihood estimate for the mean of the Gaussian distribution for class c is given by

$$\boldsymbol{\mu_c} = \frac{1}{N_c} \sum_{\substack{n=1 \\ y^n = c}}^{N} \boldsymbol{x}^{(n)}$$

which represents the mean of the observations assigned to class c.
Similarly, show that the maximum likelihood estimate for the shared covariance matrix is given by

$$\boldsymbol{\Sigma} = \sum_{c=1}^{C} \frac{N_c}{N} \boldsymbol{S_c} \qquad \text{where,} \qquad \boldsymbol{S_c} = \frac{1}{N_c} \sum_{\substack{n=1 \\ y^n = c}}^{N} (\boldsymbol{x}^{(n)} - \boldsymbol{\mu_c})(\boldsymbol{x}^{(n)} - \boldsymbol{\mu_c})^T$$

Thus $\mathbf{\Sigma}$ is given by a weighted average of the sample covariances of the data associated with each class,in which the weighting coefficients $\frac{N_c}{N}$ are the prior probabilities of the classes.