Machine Learning Exercise Sheet 05

Linear Classification

Exercise sheets consist of two parts: In-class exercises and homework. The in-class exercises will be solved and discussed during the tutorial. The homework is for you to solve at home and further engage with the lecture content. Note that the order of some exercises might have changed compared to last year's recordings.

In-class Exercises

Multi-Class Classification

Problem 1: Consider a generative classification model for C classes defined by class probabilities $\mathrm{p}(y = c) = \boldsymbol{\pi}_c$ and general class-conditional densities $\mathrm{p}(\boldsymbol{x} \mid y = c, \boldsymbol{\theta}_c)$ 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.

Linear Discriminant Analysis

Problem 2: 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

$$\mathrm{p}(\boldsymbol{x} \mid y = c, \boldsymbol{\theta}) = \mathrm{p}(\boldsymbol{x} \mid \boldsymbol{\theta}_c) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_c, \boldsymbol{\Sigma}).$$

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

$$\boldsymbol{\Sigma} = \sum_{c=1}^{C} \frac{N_c}{N} \boldsymbol{S}_c \quad \text{where} \quad \boldsymbol{S}_c = \frac{1}{N_c} \sum_{\substack{n=1 \\ y^{(n)}=c}}^{N} (\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)^{\mathrm{T}}.$$

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 N_c/N are the prior probabilities of the classes.

Homework

Linear classification

Problem 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

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

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

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

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

a) Suppose you are given an observation x. What is the name of the posterior distribution $\mathrm{p}(y \mid x)$? You only need to provide the name of the distribution (e.g., "normal", "gamma", etc.), not estimate its parameters.

b) What values of x are classified as class 1? (As usual, we assume that the classification decision is $\hat{y} = \arg\max_k \mathrm{p}(y = k \mid x)$)

Problem 4: Let $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}$ be a linearly separable dataset for 2-class classification, i.e. there exists a vector $\boldsymbol{w}$ such that $\operatorname{sign}\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}\right)$ separates the classes. Show that the maximum likelihood parameter $\boldsymbol{w}$ of a logistic regression model has $\|\boldsymbol{w}\| \to \infty$. Assume that $\boldsymbol{w}$ contains the bias term.

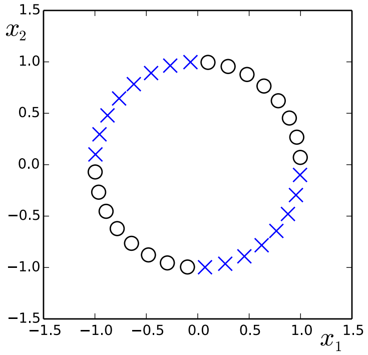
How can we modify the training process to prefer a $\boldsymbol{w}$ of finite magnitude?

Problem 5: In the 2-class case, show that a softmax of a linear function is equivalent to a sigmoid of a linear function.

Problem 6: Show that the derivative of the sigmoid function $\sigma(a) = \left(1 + e^{-a}\right)^{-1}$ can be written as

$$\frac{\partial \sigma(a)}{\partial a} = \sigma(a)\left(1 - \sigma(a)\right).$$

Problem 7: Give a basis function $\phi(x_1, x_2)$ that makes the data in the example below linearly separable (crosses in one class, circles in the other).



Naive Bayes

Problem 8: In 2-class classification the decision boundary Γ is the set of points where both classes are assigned equal probability,

$$\Gamma = \{\boldsymbol{x} \mid \mathrm{p}(y = 1 \mid \boldsymbol{x}) = \mathrm{p}(y = 0 \mid \boldsymbol{x})\}\,.$$

Show that Naive Bayes with Gaussian class likelihoods produces a quadratic decision boundary in the 2-class case, i.e. that Γ can be written with a quadratic equation of $\boldsymbol{x}$,

$$\Gamma = \left\{\boldsymbol{x} \mid \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{x} + c = 0\right\},$$

for some $\boldsymbol{A}$, $\boldsymbol{b}$ and c.

As a reminder, in Naive Bayes we assume class prior probabilities

$$\mathrm{p}(y = 0) = \boldsymbol{\pi}_0 \quad \text{and} \quad \mathrm{p}(y = 1) = \boldsymbol{\pi}_1$$

and class likelihoods

$$\mathrm{p}(\boldsymbol{x} \mid y = c) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

with per-class means $\boldsymbol{\mu}_c$ and *diagonal* (because of the feature independence) covariances $\boldsymbol{\Sigma}_c$.