Machine Learning Exercise Sheet 05

Linear Classification

1 Linear classification

Problem 1: We want to create a generative binary classification model for classifying *nonnegative* one-dimensional data. This means, that the labels are binary ($y \in \{0, 1\}$) and the samples are $x \in [0, \infty)$.

We place a uniform prior on y

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

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

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

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

a) 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.

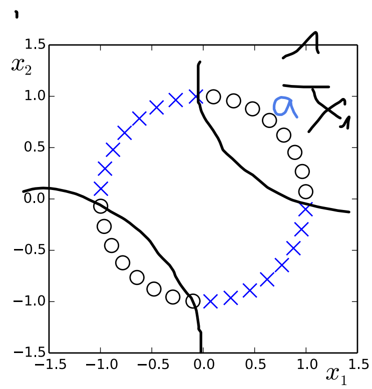
b) What values of x are classified as class 1?
(As usual, we assume that the classification decision is $y_{predicted} = \arg\max_k p(y = k \mid x)$)

Problem 2: Assume you have a linearly separable data set. What properties does the maximum likelihood solution for the decision boundary $\boldsymbol{w}$ of a linear regression model have? Assume that $\boldsymbol{w}$ includes the bias term.

What is the problem here and how do we prevent it?

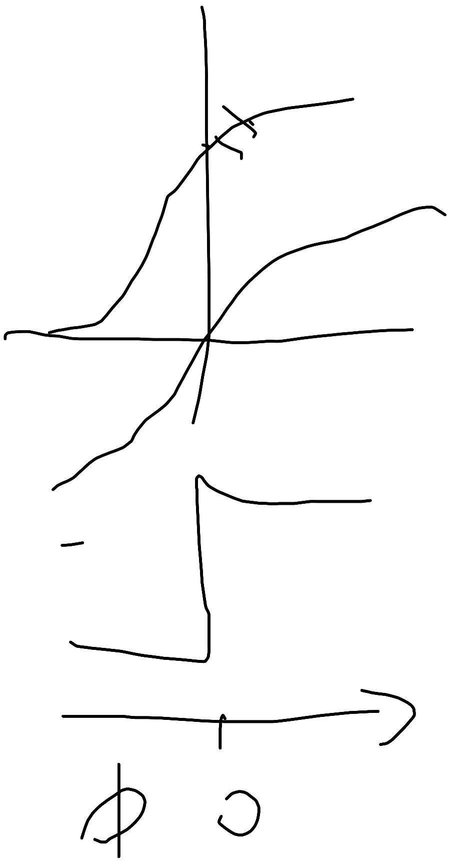
Problem 3: Show that the softmax function is equivalent to a sigmoid in the 2-class case.

Problem 4: Which basis function $\phi(x_1, x_2)$ makes the data in the example below linearly separable (crosses in one class, circles in the other)?



$(1/0.01)$
$(0.01\ 1)$

$1 = \frac{1}{a \cdot 0{,}01} \Rightarrow a = 100$



$\phi(x_1, x_2) = \sin(\underset{\pi x_1}{x_1}) \cdot \sin(\underset{\pi x_2}{x_2})$

$p(y=0 \mid x_1, x_2) = \arg\max \phi$

$p(y=x \mid x_1, x_2) = \arg\min \phi$

$\phi(x_1, x_2) = \begin{cases} 0, & \text{if } \phi > 0 \\ x & \text{else} \end{cases}$

$y = \mathrm{sig}(\phi) - 0{,}5$

In-class Exercises

2 Multi-Class Classification

Problem 5: Consider a generative classification model for C classes defined by prior class probabilities $p(y = c) = \pi_c$ and general class-conditional densities $p(\boldsymbol{x}|y = c, \boldsymbol{\theta}_c)$ where $\boldsymbol{x}$ 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 iid, show that the maximum-likelihood solution for the prior probabilities is given by

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

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

Problem 6: 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

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

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

$$\boldsymbol{\mu}_c = \frac{1}{N_c} \sum_{\{n|\boldsymbol{x}^{(n)} \in C_c\}} \boldsymbol{x}^{(n)}$$

which represents the mean of those feature vectors assigned to class C_c.

Similarly, show that the maximum likelihood solution for the shared covariance matrix is given by

$$\boldsymbol{\Sigma} = \sum_{c=1}^{C} \frac{N_c}{N} \mathbf{S}_c$$

where

$$\mathbf{S}_c = \frac{1}{N_c} \sum_{\{n|\boldsymbol{x}^{(n)} \in C_c\}} (\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)^T.$$

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