UiT
THE ARCTIC UNIVERSITY OF NORWAY

3. Parametric Methods and Bayesian Decision Theory

FYS-2021 Exercises
Department of Physics and Technology
Faculty of Science and Technology

Bayesian decision theory

Problem 1

(1a) Derive the decision boundary between two normal distributions $\mathcal{N}(\mu_0, \sigma_0), \mathcal{N}(\mu_1, \sigma_1)$.

(1b) In the univariate two-class case, explain why the Bayes' classifier minimizes the probability of misclassification when the true probability density functions are known.

Parametric methods

Problem 2

In this exercise we are going to use the bayesian classifier to classify whether a person has a low or high education based on their salary. Use the file `education-salary.csv`. The file contains two columns. The first has number 0,1 and 2 which corresponds to education levels: 0: elementary and junior high school (grunnskole), 1: High school (videregående) and 2: phD. The second column is the average monthly salary for an unspecified industry.

(2a) Make box plots and histograms for the three education levels.

Assume that the income at elementary school level (0), X_0, are normally distributed:

$$X_0 \sim \mathcal{N}(\mu_0, \sigma_0)$$

(2b) Let $\mathcal{X}_0 = \{x_1, x_2, ..., x_N\}$ denote the observed set of average monthly salaries for education level 0. Derive the expression for the log likelihood function $\mathcal{L}(\mu_0, \sigma_0 | \mathcal{X}_0)$

(2c) Derive the MLE estimates for μ_0 and σ_0.

(2d) Assuming that the incomes are normal distributed, implement a function that classifies which education level a person has based on a salary. You can use the MLE estimators as parameters in the normal distribution.

(2e) Predict, using bayesian classification, the education level someone would have if they had a monthly salary of 5000,-. Comment on the result.

Problem 3

(3a) Implement a univariate Bayesian classifier for discriminating between two normal distributions with means 0 and 1, and standard-deviations 0.5 and 0.5. Assume equal prior probabilities.

(3b) Generate 1000 samples from each of the distributions, and classify these using your implementation. Calculate the confusion matrix, and plot the histograms of the samples. Illustrate the decision boundary together with the histograms. Are all points correctly classified? Why/why not?

(3c) Generate 1000 samples from a $\mathrm{Unif}(-0.87, 0.87)$ distribution, and 1000 samples from a $\mathrm{Unif}(0.13, 1.87)$ distribution. Classify these samples using the same classifier as above. Calculate the accuracy, and compare it with the previous result. You should see a drop in accuracy, even though the mean and standard-deviation of the uniform sampling distributions are equal to the sampling distributions in (a). Why is this the case? Remember that the mean and standard deviation of a $\mathrm{Unif}(a, b)$ distribution is given by $(b + a)/2$ and $\sqrt{(b - a)^2/12}$, respectively.

Problem 4

In this problem you will use a naive Bayes' classifier to create a system for detecting whether an SMS is spam or not, based on its contents. The Bag Of Words (BOW) representation[1] of 5574 text messages is provided in `sms-spam-bow.csv`. In this file, each row represents a single text message. The first element is the label (0 = not spam, 1 = spam), and the rest of the row is the BOW-representation of the message. The raw data can be found in `sms-spam.txt`. This corpus has been collected from free or free for research sources at the Web. More info can be found at http://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection

(4a) Implement your own version of the naive Bayes' classifier. You can assume that the features are binary (Bernoulli).

(4b) Split the dataset into training- and test-sets (80 % training and 20 % test). Use the training data to estimate the parameters of the classifier, and use the test set to evaluate the performance. Report the confusion matrix and accuracy.

[1] You can read more about the bag of words representation in Chapter 5.7

Optional

Problem 5

Assume that the random vector $\boldsymbol{X} = [X_1, \dots, X_d]^{\mathrm{T}}$ has distribution

$$p(\boldsymbol{x}) = \begin{cases} C(\boldsymbol{\lambda}) e^{-\boldsymbol{\lambda}^T \boldsymbol{x}}, & \min\{x_1, \dots, x_d\} > 0 \\ 0, & \text{otherwise} \end{cases}$$

where $C(\boldsymbol{\lambda})$ is a normalizing constant, $\boldsymbol{\lambda} = [\lambda_1, \dots, \lambda_d]^{\mathrm{T}}$ is a parameter vector with strictly positive elements.

(5a) Show that, if the distribution integrates to one, we have

$$C(\boldsymbol{\lambda}) = \prod_{j=1}^{d} \lambda_j$$

(5b) Argue that the X_i's are statistically independent.

(5c) Compute the Maximum Likelihood estimator for $\boldsymbol{\lambda}$. Compare your result to the Maximum Likelihood estimator for $\lambda = \frac{1}{\beta}$ in the univariate exponential distribution.