Tutorial - Exercise sheet 4

Pattern and Speech Recognition

Introduction

In this exercise you will work with some basic probability theory and experiment with maximum likelihood estimation.
For a given set of samples x, the likelihood wrt. a probability distribution (resp. probability density) f_θ which is parametrized by some $\theta \in \mathbb{R}^n$ is defined as:

$$\mathcal{L}(\theta; x) = \prod_{x_i \in x} f_\theta(x_i)$$

Likewise, the log–likelihood is defined as:

$$\ln \mathcal{L}(\theta; x) = \sum_{x_i \in x} \ln f_\theta(x_i)$$

We are interested in finding a θ_{max} which maximizes $\mathcal{L}(\theta; x)$ or $\ln \mathcal{L}(\theta; x)$.

Probability Theory

1. (1 point) Let A, B be two events. Are the following two statements equivalent?

   $$P(A) \cdot P(B) = P(A \cap B)$$

   $$P(A \mid B) = P(A)$$

   Prove or disprove. Assume $0 > P(B) > 1$.

2. (1 point) Prove Bayes law.

3. (2 points, **Bonus Exercise**) Assume you have two random variables X and Y, normally distributed and independent. Compute $\mathrm{E}[X + Y]$ and $\mathrm{Var}[X + Y]$ (wrt. the expectation and variance of X, Y).

MLE Poisson Distribution

1. (1 point) In soccer the number of goals a team scores during a single match follows roughly a Poisson distribution. In its first 10 matches the 1. FC Saarbrücken scored

   $$x = 0, 2, 2, 1, 2, 1, 3, 1, 1, 5$$

goals. Derive the likelihood–function and plot it for $\theta \in (0, 5]$. Report which value θ_{max} maximizes the likelihood. The Poisson distribution is defined as:

$$P_\theta(k) = \frac{\theta^k \cdot e^{-\theta}}{k!}$$

for some $\theta \in \mathbb{R}_{>0}$

2. (1 points) Plot the log–likelihood. What do you notice? What can you say about the relationship between likelihood and log–likelihood?

3. (1 point, **Bonus exercise**) Derive θ_{max} analytically wrt. some x.

MLE Normal distribution

1. (1 point) Again, consider the Iris data set. Import the rows for Iris–setosa and Iris–versicolor. Omit everything except for the first column, which is our new x.

2. (1 point) One might consider that x is uniformly distributed. That is, $\theta = (a, b)$ (with $a < b$) where

   $$p_\theta(x_i) = \frac{1}{b-a} \text{ if } a \leq x_i \leq b \text{ and zero otherwise.}$$

   Explain how you would choose a, b to maximize the likelihood based on your intuition.

3. (1 point) We, however, assume that x contains samples which are induced by two (equally likely) normal distributions with different μ and σ. More precisely, we assume $\theta = (\mu_1, \sigma_1, \mu_2, \sigma_2)$ and that

   $$p_\theta(x_i) = \frac{1}{2} \cdot f_{\mu_1,\sigma_1}(x_i) + \frac{1}{2} \cdot f_{\mu_2,\sigma_2}(x_i)$$

   where $f_{\mu,\sigma}$ is the probability density of the normal distribution with mean μ and standard deviation σ. Is p_θ a probability density? Explain.

4. (1 point) Implement a function which computes the likelihood and the log–likelihood for a given θ. What might be a reason for us to prefer the log–likelihood over the likelihood?

5. (1 point) Find θ_{max}. You can use either the likelihood or the log–likelihood, justify your choice. Since we need to estimate four parameters, grid search would be infeasible in this case. However, you can use any optimization library or toolbox.

   We recommend `scipy.optimize.minimize` (python) or `fminsearch` (matlab). You may use $(6.0, 0.6, 5.0, 0.5)$ as an initial guess. Choose suitable options for optimization procedure, it should at least find one local optimum.

   Plot x together with the two normal distributions for θ_{max}. Would you have expected these results?

Submission architecture

You have to generate a **single ZIP file** respecting the following architecture:

```
tutorial1_<matriculation_nb1>_<matriculation_nb2>_<matriculation_nb3>
|
+--- source
|      |
|      +------- file 1
|      +------- file 2
|      +------- ...
+--- rapport.pdf
+--- README.txt
```

where

- **source** contains the source code of your project,
- **rapport.pdf** is the report where you present your solution with **the explanations (!)** and the plots,
- **README** which contains group member informations (name, matriculation numbers and emails) and a **clear** explanation about how to compile and run your source code

The ZIP filename has to be :

```
tutorial1_<matriculation_nb1>_<matriculation_nb2>_<matriculation_nb3>.zip
```

You have to choose between the following languages **python** or **matlab**. Other languages won't be accepted.

Some hints

We advice you to follow the following guidelines in order to avoid problems :

- Avoid building complex systems. The exercises are simple enough.
- Do not include any executables in your submission, as this will cause the e-mail server to reject it.

Grading

Send your assignment to the tutor who is responsible of your group:

- Gerrit Großmann gerritgr@gmail.com
- Sébastien Le Maguer slemaguer@coli.uni-saarland.de
- Kata Naszádi b.naszadi@gmail.com

The email subject should start with `[PSR TUTORIAL 4]`