ASSIGNMENT 1

Submission deadline: classes on **??**

Points: 28.5

The following initial assignment is intended to provide a solid foundation for the mathematics required to comprehend the course. The solutions are to be presented in to the group during the classes.

We hope that you pay attention to the lectures, have fun solving problems, and enjoy the ride.

1 Probability refresher

Problem 1. **[1p]** If you want the right answers, ask the right questions

A friend of yours has two children. Assuming that conceiving a boy is equally probable as conceiving a girl, we can assume a priori that he has both a boy and a girl (or a girl and a boy) with probability $\frac{1}{2}$, two boys with probability $\frac{1}{4}$ and two girls with $\frac{1}{4}$.

(a) What is the probability that he has a daughter, if you know that he has at least one son?

(b) What is the probability that he has a daughter, if just a while ago you have seen his son playing in the garden?

Problem 2. **[1p]** Think hard and think twice

You're a guest star in a popular TV show *Go All the Way*. You have won the episode, and it's your turn to draw the prize. There are three gates you can choose from, yet only one of them holds the grand prize (the other two are empty and leave you with empty hands). The prize is assigned to one of the gates with probability $\frac{1}{3}$.

As 1 has always been your lucky number, you have chosen gate No. 1. The host shows you the content of gate No. 2, which you haven't selected. It is empty. He offers you to switch to gate No. 3. Will switching the gate increase the probability of winning the grand prize?

Problem 3. **[4p]** Bayes' theorem

Bayes' theorem allows to compute probabilities using conditional probabilities:

$$p(A, B) = p(A|B)p(B) = p(B|A)p(A) \tag{1}$$

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} \tag{2}$$

Bayes' theorem is a basis of the naïve Bayes classifier, which is often used for classification of text documents (e.g. as spam and non-spam). The naïve Bayes model assumes, that the observed document has been created in the following way: first, document class C has been drawn at random with probability $p(C)$. Then, a sequence of words $w_1, \ldots, w_n$ has been drawn at random (with replacement), each independent from the others, with probability $p(w_i|C)$. With this model, using Bayes' theorem, we can compute the probability that, knowing the words, the document comes from class C.

Solve the following using Bayes' theorem.

(a) **[2p]** There are two boxes on the table: box #1 holds two black balls and eight red ones, box #2 holds 5 black ones and 5 red ones. We pick a box at random (with equal probabilities), and then a ball from that box. What is the probability, that the ball came from box #1 if we happened to pick a red ball?

(b) **[2p]** The government has started a preventive program of mandatory tests for the Ebola virus. Mass testing method is imprecise, yielding 1% of false positives (healthy, but the test indicates the virus) and 1% of false negatives (having the virus but healthy according to test results). As Ebola is rather infrequent, lets assume that it occurs in one in a million people in Europe. What is the probability, that a random European, who has been tested positive for Ebola virus, is indeed a carrier?

Suppose we have an additional information, that the person has just arrived from a country where one in a thousand people is a carrier. How much will be the increase in probability?

How accurate should be the test, for a 80% probability of true positive in a European?

Problem 4. **[4p]** Naive Bayes classifier

A simple classifier can be devised using Bayes' theorem.

There are classes $C_1, C_2, \ldots, C_k$. Each observation is of exactly one class. For an observation X and the class C, we can apply Bayes' theorem to the probability of X's class being C:

$$p(C|X) = \frac{p(X|C)p(C)}{p(X)} \tag{3}$$

Furthermore, if X can be treated as a sequence of smaller observations $x_1 x_2 \ldots x_n$, then, using the fact that

$$(\forall A, B)\ p(A \cap B) = p(A|B)p(B), \tag{4}$$

we can re-formulate Equation 3 as

$$p(C|X) = p(x_n|x_{n-1} \ldots x_1, C) \ldots p(x_1|C)\frac{p(C)}{p(X)}. \tag{5}$$

In the Naive Bayes classifier we assume that the observations x_i are not correlated with their predecessors (this assumption is indeed naive and rarely holds, thus the name, although this classifier is quite practical nonetheless). Under this assumption,

$$p(C|X) = (\prod_{i=1}^{n} p(x_i|C))\frac{p(C)}{p(X)}. \tag{6}$$

The probabilities $p(C_1|X), p(C_2|X), \ldots, p(C_k|X)$ for classes $C_1, C_2, \ldots, C_k$ can be found.

The class to which the highest probability corresponds is considered the class of the observation. Each probability has $p(X)$ in the denominator. It does not affect the end result (as it is a scaling factor which occurs in all probabilities), and thus is omitted, instead using the equation

$$p(X|C_i) = (\prod_{i=1}^{n} p(x_i|C_i)) * p(C_i) \tag{7}$$

for each $1 \leq i \leq k$.

The factor $p(C_i)$ is called **a priori probability**. It denotes the probability that a randomly picked observation is of class C_i. If no such information is available, we can assume that $(\forall 1 \leq i \leq k)\ p(C_i) = \frac{1}{k}$. The model parameters are then the probabilities $p(x_i|C_j)$.

We will try to mimic the language-guessing feature of Google Translate (https://translate.google.com/), although on a much smaller scale. We are given an input which is a lower-case sequence of characters (such as **some people like pineapple on pizza**), and we determine whether the sequence's language is English (E), Polish (P) or Spanish (S). We will treat each character as a separate observation. We only know the probabilities of vowels **a**, **e**, **i**, **o**, **u**, **y**, so we will treat all other characters as one class. The numbers are taken from https://en.wikipedia.org/wiki/Letter_frequency#Relative_frequencies_of_letters_in_other_languages.

$p(\downarrow \mid \rightarrow)$	E	P	S
a	0.08167	0.10503	0.11525
e	0.12702	0.07352	0.12181
i	0.06966	0.08328	0.06247
o	0.07507	0.02445	0.08683
u	0.02758	0.02062	0.02927
y	0.01974	0.03206	0.01008
other	0.59926	0.66104	0.57429

(a) **[2p]** What is the language of the following phrases, according to the classifier?

i. bull,

ii. burro,

iii. kurczak,

iv. pollo,

v. litwo, ojczyzno moja, ty jesteś jak zdrowie,

vi. dinero,

vii. mama just killed a man put a gun against his head,

viii. maradona es mas grande que pele.

(b) **[1p]** Let us assume that the a priori probabilities are $p(E) = 0.5$, $p(P) = 0.2$, $p(S) = 0.3$. How will the results change?

(c) **[1p]** Try to fool the classifier! Create a phrase which is misclassified.

2 Maximum likelihood estimation

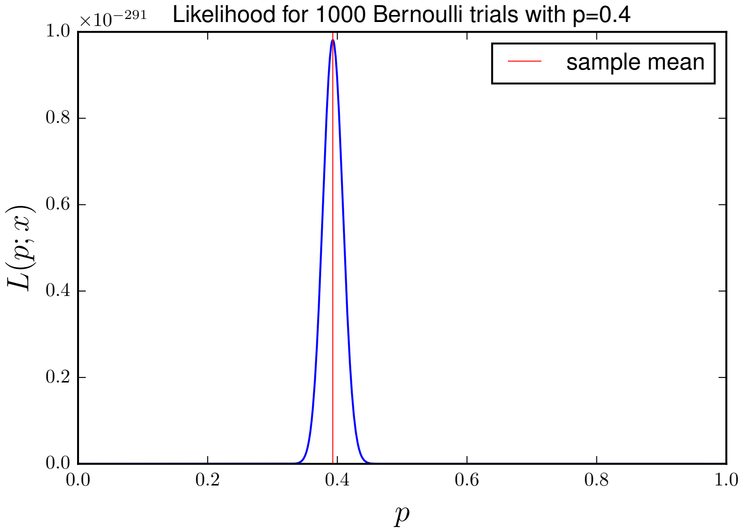
Problem 5. **[3p]** Given observations $x_1, \ldots, x_n$ coming from a certain distribution, prove that MLE of a particular parameter of that distribution is equal to the sample mean $\frac{1}{n}\sum_{i=1}^{n} x_i$:

(a) **[1p]** Bernoulli distribution with success probability p and MLE $\hat{p}$,

(b) **[1p]** Gaussian distribution $\mathcal{N}(\mu, \sigma)$ and MLE $\hat{\mu}$,

(c) **[1p]** Poisson distribution $Pois(\lambda)$ and MLE $\hat{\lambda}$.

Problem 6. **[2p]** Draw n random samples from the Bernoulli distribution. Plot sample mean and likelihood as a function of p. Experiment with $n \in \{50, 100, 500, 1000\}$ and $p \in \{0.4, 0.5\}$. You should obtain a plot like the one below:



Tip: To make a plot, evaluate $L(p; x)$ in a finite set of points, for instance:
`p = linspace(0.0, 1.0, 1000)`

Problem 7. [**3p**] There is a game called **rock, paper, scissors**. Klapaucius has built a robot which plays the game (it displays **R**, **P** or **S** – the first letter of the appropriate symbol – on a monitor). The robot has won the world championship, however, its memory drive – including its configuration – has been irreparably damaged due to an overly excessive celebration, including the consumption of unhealthy amounts of grease.

The robot followed a rather simple algorithm. The choice of the symbol was randomised, and it depended only on the previous symbol. The algorithm has nine parameters – $\{p(a|b) : a, b \in \{R, P, S\}\}$ – denoting the probability of showing symbol a if the most recently shown symbol was b. Klapaucius has a terrible memory for numbers and he does not remember the configuration. Fortunately, he has got hold of the video footage from which he was able to decipher the sequence of symbols shown. The sequence is as follows:

PPRSSRSPPRSPRRSPPPSSPRSPSPSRSP

(a) [**1p**] Find the most likely configuration, given that the first symbol has been chosen totally at random.

(b) [**1p**] Find the most likely configuration, given that the first symbol has been chosen totally at random. The robot's random number generator is actually a monkey rolling a six-sided die.

(c) [**1p**] Trurl is the runner-up. He decided to copy Klapaucius's robot. He decided, however, that there is no need for the probabilities to be based on the previous symbol. His configuration had three parameters – $p(R)$, $p(P)$, $p(S)$. He set them according to the same footage. Find the parameters. Simulate a million games between the two robots. One game lasts until one side has won three rounds. Which robot wins more frequently?

3 Gradients

Problem 8. [**4p**] Find the following functions' gradients over $[\mathbf{x}, \mathbf{y}, \mathbf{x}]^\mathbf{T}$:

(a) [**0.5p**] $f_1([x, y, x]^T) = x + y$,

(b) [**0.5p**] $f_2([x, y, x]^T) = xy$,

(c) **[0.5p]** $f_3([x,y,x]^T) = x^2y^2$,

(d) **[0.5p]** $f_4([x,y,x]^T) = (x+y)^2$,

(e) **[0.5p]** $f_5([x,y,x]^T) = x^4 + x^2yz + xy^2z + z^4$,

(f) **[0.5p]** $f_6([x,y,x]^T) = e^{x+2y}$,

(g) **[0.5p]** $f_7([x,y,x]^T) = \frac{1}{xy^2}$,

(h) **[0.5p]** $f_8([x,y,x]^T) = \tanh(ax+by+c)$.

Problem 9. [1p] Find the following functions' derivatives over $\mathbf{x}$:

(a) **[0.5p]** $\tanh(\mathbf{x})$,

(b) **[0.5p]** $\sigma(\mathbf{x}) = \frac{1}{1+e^{-\mathbf{x}}}$.

Problem 10. [3.5p] Find the following functions' gradients over the vector $\mathbf{x}$ ($\mathbf{x}, \mathbf{b} \in \mathbb{R}^n$, $\mathbf{W} \in \mathbb{R}^{n\times n}$):

(a) **[0.5p]** $\mathbf{Wx} + \mathbf{b}$,

(b) **[1p]** $\mathbf{x}^T\mathbf{Wx} + \mathbf{b}$,

(c) **[1p]** $\sigma(\mathbf{Wx} + \mathbf{b})$,

(d) **[1p]** $S(\mathbf{x})$, where S is the softmax function (https://en.wikipedia.org/wiki/Softmax_function).

4 Numerical properties

Problem 11. [2p] Neural networks, as well as the broader subject, data analysis, is often computation-heavy. The problems and limitations mensioned during the Numerical Analysis course often occur here.

(a) **[1p]** Why do we want to avoid operations like $(\frac{1}{2})^{10000}$? What is the result of $\frac{0.5^{10000}}{0.5^{10001}}$?

(b) **[1p]** How can a logarithm be used to avoid some numerical problems (hint: look up *log-likelihood*)?