# SSI Pattern Recognition

Homework 1

Computer Vision

Désiré Sidibé
dro-desire.sidibe@u-bourgogne.fr

## Linear Regression

### I. Introduction

In this homework, we will review some important tools of matrix algebra and probability theory, and we will implement linear regression and study the effect of regularization.
You shall submit a clearly written and commented report.

### II. Matrix algebra

In the following, $\mathbf{a}$ and $\mathbf{b}$ are vectors and $\mathbf{A}$ is a matrix.
Prove the following useful results:

1. $$\frac{\partial(\mathbf{b}^T\mathbf{a})}{\partial\mathbf{a}} = \mathbf{b}^T$$
2. $$\frac{\partial(\mathbf{A}\mathbf{a})}{\partial\mathbf{a}} = \mathbf{A}$$
3. $$\frac{\partial(\mathbf{a}^T\mathbf{A}\mathbf{a})}{\partial\mathbf{a}} = \mathbf{a}^T(\mathbf{A} + \mathbf{A}^T)$$
4. $$\frac{\partial}{\partial\mathbf{A}}(\text{trace}(\mathbf{B}\mathbf{A})) = \mathbf{B}$$
5. $$\text{trace}(ABC) = \text{trace}(CAB) = \text{trace}(BCA)$$

**Note**: these are important and useful results that you need to know.

### III. Maximum Likelihhod (ML) estimate

Suppose you are given a set of $N$ observations of a scalar variable $x$, $\mathbf{x} = (x_1, x_2, \ldots, x_N)^T$. We assume that the observations are drawn independently from a Gaussian distribution whose mean $\mu$ and variance $\sigma^2$ are unkwon. We would like to determine these parameters from the data set.
The likelihood of the data given the parameters is given by

$$p(\mathbf{x}|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2).$$

1. Explain why the likelihood function can be written as a product of Gaussians?
2. ML aims to find the values of the parameters that maximize the likelihood function. Since $\ln(x)$ is a monotonically increasing function of $x$, we can instead maximize the log of the likelihood.
   (a) Compute the log likelihood function: $\ln p(\mathbf{x}|\mu, \sigma^2)$.
   (b) By taking the partial derivatives of the log likelihood with respect to $\mu$ and to $\sigma^2$, find the ML solutions $\mu_{ML}$ and $\sigma^2_{ML}$.

**IV. Linear Regression with regularization**

You are given a file `hw1training.txt` which contains training data that will be used for linear regression. The first column contains the input feature $x$ and the second column contains the output value $y = f(x)$ for some unknown function $f$.
We will assume a polynomial function of the form

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j.$$

1. Write a function `w = linearRegress(X, y, M)` that estimates the parameters $\mathbf{w}$ of the model given $X$, $y$ and the degree $M$ of the polynomial.
2. Try different values of $M$ and show how well the obtained model fits the data?

In order to select a "good" model, i.e. a good value for $M$, we will use another dataset provided in the file `hw1test.txt`. As shown in class, for each model, we can compute both the training error and the test error.

1. Compute and plot the training and test error for different values of $M$ in a single figure.
   Note: you have to use the root-mean-square (RMS) error defined by $E_{RMS} = \sqrt{2E(\mathbf{w}^*)/N}$, where $E(\mathbf{w}^*)$ is the empirical error computed with the obatined optimal value $\mathbf{w}^*$, and $N$ is the number of data samples:
   $$E(\mathbf{w}^*) = \frac{1}{2}\sum_{i=1}^{N}(\mathbf{w}^{*T}\phi(\mathbf{x}_i) - y_i)^2.$$
2. What is a good model for this data? Justify your answer.

We now want to fit a model of order $M = 10$ to the data.

1. Explain how we can perform this task using regularization.
2. Find a good regularization parameter and apply it to the data. Display and analyze your results.

**V. Additional (optional):** Regularization = MAP

We consider a regression problem given a training data set $\mathbf{x} = (x_1, x_2, \ldots, x_N)^T$ with corresponding target values $\mathbf{t} = (t_1, t_2, \ldots, t_N)^T$.

Let assume that given the value of $x$, the corresponding target value of $t$ has a Gaussian distribution with mean equal to $y(x, \mathbf{w})$. Thus

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}),$$

where $\beta^{-1}$ is the precision parameter equal to the inverse variance of the distribution.
As in the case above, we want to use the training data $\{\mathbf{x}, \mathbf{t}\}$ to determine the values of the unknown parameters $\mathbf{w}$ and $\beta$. If the data is i.i.d., then the likelihood function is given by

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|y(x_n, \mathbf{w}), \beta^{-1}).$$

1. Show that maximizing the likelihood function, to find $\mathbf{w}$, is equivalent to minimizing the *sum-of-squares* error function as defined in class.

   Note: the sum-of-squares error function is given by $\sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2$.

2. We can also introduce a prior information over the polynomial coefficients $\mathbf{w}$ as follows

   $$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}),$$

   where $\alpha^{-1}$ is the precision of the distribution.

   Using Bayes theorem, the posterior distribution for $\mathbf{w}$ is proportional to the product of the prior distribution and the likelihood function:

   $$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha).$$

   (a) Show that maximizing this posterior distribution, to find $\mathbf{w}$, is equivalent to minimizing the regularized sum-of-squares error function.

   (b) Give the value of the regularized parameter $\lambda$ in this case.

   This technique is called *maximum a posteriori* (MAP).