Prof. Dr. Volker Roth
volker.roth@unibas.ch
6^{th} floor

Sebastian Keller
sebastianmathias.keller@unibas.ch
1^{st} floor

Departement Mathematik und Informatik
Spiegelgasse 1
4051 Basel

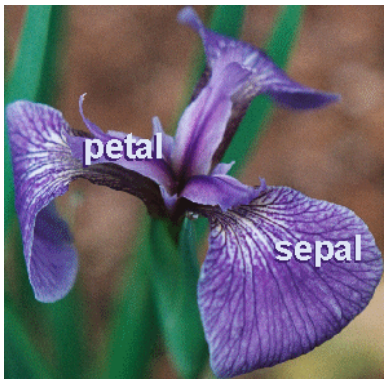
University of Basel

Solutions for exercises marked with a '⋆' will be made available online, usually in the following week. Only if needed will these exercises be discussed during the tutorial. All other exercises are to be prepared at home and presented by the participants during the tutorial session.

Exercise 7

7.1 Discriminant Analysis

In 1936, Sir Ronald A. Fisher introduced the Iris flower data set as an example of discriminant analysis. It consists of 50 samples of 3 Iris flower species (Iris setosa, Iris virginica and Iris versicolor). 4 features were measured from each sample, being the length and the width of sepal and petal. Based on this, Fisher developed a linear discriminant model to determine which species a sample belongs to.



The data for classes *setosa* and *virginica* are provided by files `setosa.csv` and `virginica.csv`, where each row corresponds to one sample. In order to simplify the problem, we restrict ourselves to the first 2 features only. The goal is to separate both classes by Fisher's Discriminant Analysis (the least-squares version utilizing SVD). Keep in mind, the data must be centered such that the sample mean satisfies

$$\boldsymbol{m} = \sum_{i=1}^{n} \boldsymbol{x}_i = \mathbf{0},$$

where $\boldsymbol{x}_i$ is a sample and n is their total number.

Write a Python script to solve the following problem:

1. Read in both data matrices from files, rows are samples and columns are features.
2. Get n, n_1 and n_2, i.e. the total number of samples and the number of samples per class.
3. Construct the matrix A containing the samples from both classes. Remember: the columns of matrix A must be centered.
4. The problem setting is: $A\mathbf{c} = \mathbf{b}$.
5. The size of $\mathbf{b}$ is n, what is its content?
6. Solve for $\mathbf{c}$, do some appropriate plotting.

7.2 PCA from scratch*

For the `hald.csv` data set from last weeks' exercise sheet we now want to program our own implementation of PCA from scratch, i.e. without using Scikit-learn.

Write a Python script to solve the following problem:

- Calculate the empirical covariance matrix.
- Use the eigenvalue decomposition (EVD) to obtain the rotation matrix $\mathbf{U}$.
- Is it important whether you perform the eigendecomposition of the covariance matrix before or after standardizing the data?
- Generate appropriate visualizations in order to check whether your implementation performs correctly.

7.3 Secondary Structure Prediction*

The problem of predicting helix structures in protein sequences was introduced in the lecture where we derived an algorithmic variant of classification by least squares. Here, the advantage was that only a "small" matrix $A^\top A$, vector $A^\top \boldsymbol{b}$ and scalar $\boldsymbol{b}^\top \boldsymbol{b}$ were needed to solve large systems of the form $A\boldsymbol{x} = \boldsymbol{b}$ for $\boldsymbol{x}$.

As the system is overdetermined, our goal is to find the solution vector $\boldsymbol{x}$ which minimizes $\|\boldsymbol{r}\|^2 = \|A\boldsymbol{x} - \boldsymbol{b}\|^2$. We will analyze how $\|\boldsymbol{r}\|^2$ and $\boldsymbol{x}^\top \boldsymbol{x}$ change when zeroing a varying number of singular values.

Please make sure to read the lecture slides! Given are the two files `AtA.csv` and `Atb.csv` containing the result of the calculations $\mathbf{A^t A}$ and $\mathbf{A^t b}$ and the additional information $\mathbf{b^t b} = 48$. Write a Python script to solve the following problem:

- For solving $\mathbf{Ax} = \mathbf{b}$ without using $\mathbf{A}$ or $\mathbf{b}$ but using $\mathbf{A^t A}$ and $\mathbf{A^t b}$ instead, we need the singular values $\mathbf{S}$ and the matrix $\mathbf{V}$ (for computing $\mathbf{x} = \mathbf{Vz}$). See lecture slides page 155.
- Find $\mathbf{z}$ which minimizes $\|\boldsymbol{S}\boldsymbol{z} - \boldsymbol{c}\|^2$
- Compute $\mathbf{x}$ that corresponds to $\mathbf{z}$, i.e. $\mathbf{x} = \mathbf{Vz}$. Calculate $\mathbf{x^t x}$.
- Calculate the error for $\mathbf{x}$ (using $\mathbf{A^t A}$, $\mathbf{A^t b}$ and $\mathbf{b^t b}$), i.e. calculate $\mathbf{r^t r} = \mathbf{x^t\, A^t A\, x} - \mathbf{2x^t\, A^t b} + \mathbf{b^t b}$.
- In two different plots, plot $\mathbf{r^t r}$ and $\mathbf{x^t x}$ as a function of the number of singular values used. There is a total of 21 non-zero values in the matrix $\mathbf{S}$.