$\mathrm{Cov}(X,Y) < \sqrt{\mathrm{Var}(X)\ \mathrm{Var}(Y)}$

Cognitive Algorithms - Exercise Sheet 2

Linear Discriminant Analysis (LDA)

Department of Machine Learning - TU Berlin

Disclaimer

For each exercise, but particularly for exercises involving calculations:

Show your work or you will not receive (full) credit!

Furthermore: Exercises marked with an asterisk * are not required and don't contribute towards the credit you receive, but you are welcome to do them because they are fun.

Introduction

Remember the definition of a covariance matrix.

Definition 0.1 (Covariance Matrix)**.** *The covariance matrix* $\Sigma_{\mathbf{v}} \in \mathbb{R}^{d\times d}$ *of a vector of random variables* $\mathbf{v} \in \mathbb{R}^d$ *is defined as:*

$$\forall i,j \in \{1,\dots,d\} : [\Sigma_{\mathbf{v}}]_{i,j} = \mathrm{Cov}(v_i, v_j) = \mathbb{E}[(v_i - \mathbb{E}[v_j])(v_j - \mathbb{E}[v_i])].$$

Task 1 - Covariance Basics [5 points]

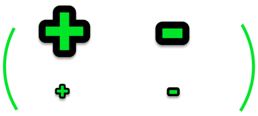

Figure 1: The first row of this matrix has elements with large values. The second row has elements with small values. The first column is positive; the second column is negative.

For questions 1, 2, and 3:
Consider a set of two random variables $\mathbf{X} = \{X_1, X_2\}$, such that each is real-valued and normally distributed with mean 0. Furthermore, X_1 has greater variance than X_2. The covariance $\mathrm{Cov}(X_1, X_2)$ is either

a) positive
b) zero
c) negative

For a, b, and c, answer the following three questions

1. What is the dimensionality of the covariance matrix $\Sigma_{\mathbf{X}}$ and why? **[0.5 points]**

2. Give $\Sigma_{\mathbf{X}}$ using the following notation: **[1.0 points]**
Note: For every positive number, write "+". For every negative number, write "-". For 0, write "0". The sizes of the +'s and -'s should be relative to their absolute value. Figure 1 illustrates the correct use of this notation.

3. Plot data with covariance described by $\Sigma_{\mathbf{X}}$ (don't worry about units). **[3 x 1 points]**

4. Which property is represented by the diagonal elements of all covariance matrices? **[0.5 points]**

Task 2: Empirical Covariance Matrix [9 points]

1. We have seen in the lecture that, for a centered dataset $X \in \mathbb{R}^{d \times n}$, the formula for the empirical covariance matrix on centered data (i.e. with $\boldsymbol{\mu}_X = \mathbf{0}$) is:

$$\begin{aligned}\hat{\Sigma} &= \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k \mathbf{x}_k^T \\ &= \frac{1}{n} X X^T.\end{aligned}$$

   (a) For some datapoint $\mathbf{x}_k$, what is the dimension of $\mathbf{x}_k \mathbf{x}_k^T$? **[0.5 points]**

   (b) Consider the entry $[\hat{\Sigma}]_{0,0}$. How is this calculated? Using your own words, explain what this sum represents. **[2 points]**

   (c) Why does the data have to be centered for the above formula to produce the empirical covariance matrix? Give the formula for non-centered data. **[1 point]**

2. Consider a data set with four data points

$$X = \begin{bmatrix} -1 & -1 & 1 & 1 \\ -1 & 0 & 0 & 1 \end{bmatrix}.$$

Compute the empirical covariance matrix $\hat{\Sigma}$. **[1 point]**

3. Covariance matrices have two important properties: They are symmetric and positive semi-definite. This also true for empirical covariance matrix estimates.

Remember, a matrix A is positive semi-definite (PSD) if, for all $\mathbf{v} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$, the following inequality holds

$$\mathbf{v}^T A \mathbf{v} \geq 0.$$

   (a) Prove that the empirical covariance matrix is always positive semi-definite **[3 points]**
   Hint: Substitute the formula for the empirical covariance matrix into the definition for PSD

   (b) Prove that the empirical covariance matrix is always symmetric **[1.5 points]**
   No hint for this one :)

Task 3 - PSD: Pretty Sweet, Dude! [*]

We care in particular about PSD matrices because all covariance matrices are PSD. One interesting property about PSD matrices is that all of their eigenvalues are non-negative. In particular it holds that all real, symmetric matrices A with non-negative eigenvalues are PSD, and vice versa. Prove this statement in both directions!

Hint 1: For the direction

$$\text{PSD} \Rightarrow \text{Eigenvalues are positive}$$

start with $A\mathbf{v} = \lambda\mathbf{v}$, for some eigenvector $\mathbf{v}$ of A.

Hint 2: For the direction

$$\text{Eigenvalues are positive} \Rightarrow \text{PSD}$$

consider that for all real symmetric matrices A, they can be decomposed using an eigenvalue decomposition, such that $A = Q\Lambda Q^T$, where Q is an orthogonal matrix and Λ contains the eigenvalues of A

Task 4 - NCC vs. LDA [6 points]

Last lecture we looked at NCC. This lecture we looked at LDA. These models are closely related, as you saw in the lecture. They are both linear classifiers, which means they both work using a weight vector $\mathbf{w}$ and a bias β, and classify points $\mathbf{x}_i$ using the expression: $\mathbf{w}^T\mathbf{x}_i - \beta \geq 0$
Recall the definition for LDA given in the lecture, given a dataset $X \in \mathbb{R}^{d\times n}$ and class means $\bar{\mathbf{x}}_+, \bar{\mathbf{x}}_- \in \mathbb{R}^d$

$$\begin{aligned}\mathbf{w} &= \bar{S}^{-1}(\bar{\mathbf{x}}_+ - \bar{\mathbf{x}}_-)\\ \beta &= \frac{1}{2}\mathbf{w}^T\left(\bar{\mathbf{x}}_+ + \bar{\mathbf{x}}_-\right)\left[+\log\left(\frac{n_-}{n_+}\right)\right].\end{aligned}$$

vs for NCC

$$\begin{aligned}\mathbf{w} &= \bar{\mathbf{x}}_+ - \bar{\mathbf{x}}_-\\ \beta &= \frac{1}{2}\mathbf{w}^T\left(\bar{\mathbf{x}}_+ + \bar{\mathbf{x}}_-\right)\end{aligned}$$

LDA performs best when the features of each class are Gaussian distributed and the covariance matrices for each class are *equal*.

1. What's the point of the value in the square brackets? Consider its value if class $-$ is much larger/smaller than class $+$, or they are of equal size. **[2 points]**
2. LDA and NCC are equal given a specific condition of the covariance matrices of each class. Prove that they are equal in this case. **[*]**
   Hint: Consider the difference between the LDA and NCC formulas for the weight vector
3. Consider data in two classes $+$ and $-$ of equal size. They have class means $\bar{\mathbf{x}}_- = \begin{pmatrix}0\\1\end{pmatrix}$ and $\bar{\mathbf{x}}_+ = \begin{pmatrix}-2\\0\end{pmatrix}$, and empirical covariance matrices $S_- = \begin{pmatrix}1 & -0.99\\-0.99 & 1\end{pmatrix}$ and $S_+ = \begin{pmatrix}1 & -0.99\\-0.99 & 1\end{pmatrix}$.
   - Sketch how this dataset might look. **[1 point]**
   - Calculate $\mathbf{w}$ and β for LDA and NCC **[1 point]**
     Hint: for LDA, use a calculator!
   - Plot the decision boundaries for LDA and NCC **[1 point]**
   - Describe why LDA performed better **[1 point]**

Task 5 - Whitening [*]

Whitening is a linear transformation of a data set. Its purpose is to decorrelate data and set the variance to one, featurewise. Thus, when whitening is applied, the covariance matrix is the identiy matrix afterwards. For many algorithms, whitening is a useful preprocessing step.

Suppose we have n empirical data points $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$ with zero mean. The covariance matrix is $\Sigma_X = \frac{1}{n} X X^T$. It can be rewritten as $\Sigma_X = U \Lambda U^T$. This is the eigendecomposition[1] into an orthogonal matrix U with eigenvectors in the columns and a diagonal matrix Λ with the eigenvalues on the diagonal. Remember that for orthogonal matrices U the inverse is the transposed and thus $U^T U = U U^T = I$.

In this exercise we use the whitening operation as the mapping

$$\mathbf{z}_k := P^T \mathbf{x}_k = U \Lambda^{-\frac{1}{2}} U^T \mathbf{x}_k = \Sigma_X^{-\frac{1}{2}} \mathbf{x}_k$$

To get a geometrical intuition take a look at figures 2 and 3. Note that there are different ways to define whitening. For example one could also skip the last rotation by the matrix U visualized in figure 2.

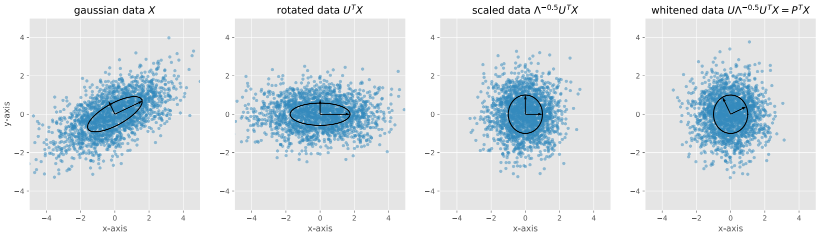

Figure 2: The whitening process: the different matrices are applied one after another.

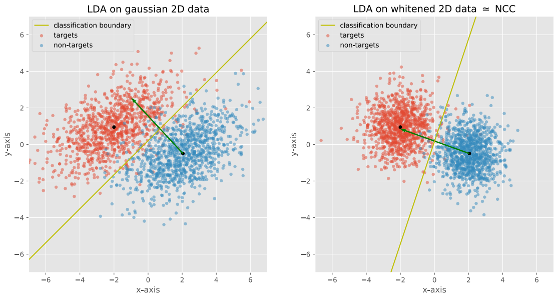

Figure 3: LDA applied to 2D gaussian toy data with and without whitening

[1] https://en.wikipedia.org/wiki/Eigendecomposition_of_a_matrix

1. Show that P is a symmetric matrix.
2. Show that
$$\Sigma_X^{\frac{1}{2}} = U\Lambda^{\frac{1}{2}}U^T$$
is a valid square root of a positive definite matrix.
3. Show that P is a valid inverse of the square root of a positive definite matrix.
4. Show that the covariance of the whitened data $P^T X$ is the identity matrix. Keep in mind that we deal with data with zero mean.
5. Proof that classification with LDA is equivalent to classification with NCC of the whitened data.