LEARNING ALGORITHMS AND SYSTEMS LABORATORY
FACULTE DES SCIENCES ET TECHNIQUES DE L'INGENIEUR
SCHOOL OF ENGINEERING / SWISS INSTITUTE OF TECHNOLOGY

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Lecturer:
Prof. ***Aude Billard***
Office: *ME A3 464*
direct: *+41-21-693 54 64*
secretary:+41-21-693 09 39
fax: *+41-21-693 78 50*
E-mail: aude.billard@epfl.ch
Web: http://lasa.epfl.ch/

Assistant:
Dr. **Basilio Noris**
Office: *ME A3 395*
direct: +41-21-693 7824
E-mail: basilio.noris@epfl.ch

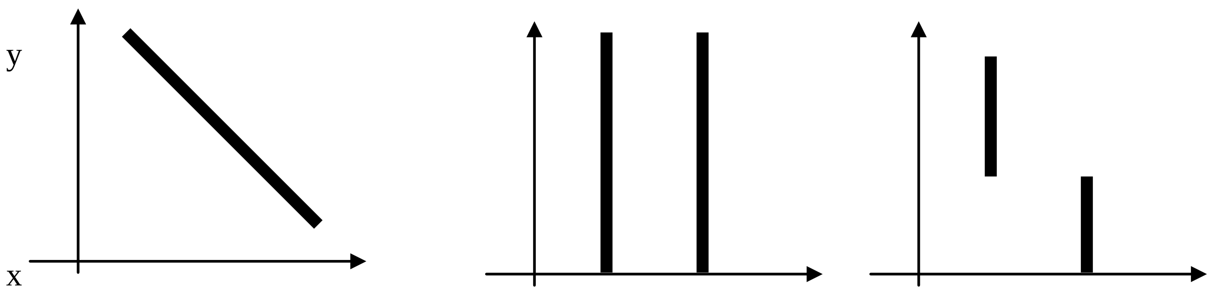
Applied Machine Learning

Exercises - I

WINTER 2011-2012

Exercise 1: Principal Component Analysis

a) For which of the following distributions are the two variates x and y statistically independent?



Solution:
Left: x and y are linearly correlated

Middle: x and y are independent, as knowing the value of x does not tell you anything about the value of y and vice-versa.

Right: x and y are correlated by could be explained by two locally independent joint distributions.

b) Imagine that you have run PCA on data gathered from one of the questionnaires gathered by the car manufacturer in which 10’000 people gave their age, gender, country of residence and the car model they purchased. How would you interpret the results if:

i) You find that all N eigenvectors (N=dimensionality of the dataset) covers the same % of the data variance.

ii) 1 single eigenvector covers 99% of the data variance (5%).

Solution:

The percentage of data variance covered by each eigenvector is directly given by its eigenvalue.

Each projection $X_i', \ i=1...N,$ of X is given by $X_i' = \left(\left(e^i\right)^T X\right)e^i$

The percentage of the dataset covered by each projection is hence given by:

$$\frac{\left\|X^T e^i\right\|}{\left\|\sum_j X^T e^j\right\|}.$$

The numerator can be expanded as follows:

$$\left(X^T e^i\right)^T X^T e^i = \left(e^i\right)^T XX^T e^i = \left(e^i\right)^T \lambda_i e^i = \lambda_i.$$

The same can be done with the numerator. Using also the fact that for all pairs $(i,j), \ i \neq j \ \left(e^i\right)^T e^j = 0,$:

$$\frac{\left\|X^T e^i\right\|}{\left\|\sum_j X^T e^j\right\|} = \frac{\lambda_i}{\sum_j \lambda_j}.$$

i) We have $\frac{\lambda_i}{\sum_j \lambda_j} = cst \ \forall i = 1...N$, i.e $\lambda_1 = \lambda_2 = \lambda_3 = ... = \lambda_N$. .

One may think that knowing the person's age, gender and country of residence did not convey enough information to determine the choice of model by a given person.

This would be true only if the 10000 people interrogated in the survey covered well the distribution of possible age, gender and country of residence.

One may thus either conclude that the survey was not well run (poor statistics or poor choice of question) or that the car manufacturer must continue producing all the current model as there is not one model more popular than another in any one country.

ii) We have $\frac{\lambda_1}{\sum_i \lambda_i} = 0.99$

This means that all factors in the analysis were linearly correlated. This indicates that the survey was not representative, as gender, age and country should have had no correlation with one another.

Exercise 2: Statistical Independence and uncorrelatedness

i) Consider two variables $x = [x_1 \ x_2] \in \mathbb{R}^2$ and $y \in \mathbb{R}$ with joint distribution $\mathrm{p}_{x,y}(x, y)$. Explain which of the following two joint distributions are statistically independent and why:

$\mathrm{p}_{x,y}(x, y) = (1 - x_1)(x_1 - y)$

$\mathrm{p}_{x,y}(x, y) = (x_1 + 3x_2)\, y$

ii) If x_1 and x_2 are two uncorrelated variables, show that if g and h are two linear functions, then $y_1 = g(x_1)$ and $y_2 = h(x_2)$ are still uncorrelated.
Show that if g and h are non linear, integrable function, y_1 and y_2 are uncorrelated if x_1 and x_2 are statistically independent.

ii) Show that when a N-dim. set of data points X is projected onto the eigenvectors $\mathrm{V}=[\mathrm{v}_1 ... \mathrm{v}_N]$ of its covariance matrix $\mathrm{C}=\mathrm{XX}^T$, the covariance matrix YY^T of the projected data Y is diagonal and hence that, in the space of the eigenvector decomposition, the distribution of X is uncorrelated.

Solution:

i) The answer is no for the first example as $p_{x,y}(x, y)$ is not factorizable into two separate pdf that depend solely on x and y respectively.
The answer is yes for the second examples as $p_{x,y}(x, y)$ is factorized into $p_x(x) = (x_1 + 3x_2)$ and $p_y(y) = y$. Note that x_1, x_2 are not independent.

---

ii) $E\{y_1, y_2\} = E\{g(x_1), h(x_2)\}$

Since $g(x_1)$ and $h(x_2)$ are two linear function, they can be expressed as through a matricial product:

$E\{y_1, y_2\} = E\{A_1 x_1, A_2 x_2\}$

Using the factorization of Expectation, we get

$E\{y_1, y_2\} = A_1 E\{x_1, A_2 x_2\} = A_1 A_2 E\{x_1, x_2\};$

Since x_1, x_2 are uncorrelated

$A_1 A_2 E\{x_1, x_2\} = A_1 A_2 E\{x_1\} E\{x_2\}$

Which again we can write as:

$E\{A_1 x_1\} E\{A_2 x_2\} = E\{y_1\} E\{y_2\}.$

ii)

if x_1 and x_2 are statistically independent, $p(x_1, x_2) = p(x_1)\,p(x_2)$,

$$E\{y_1 \cdot y_2\} = E\{g(x_1) \cdot h(x_2)\}$$
$$= \iint g(x_1)\,h(x_2)\,p(x_1, x_2)\,dx_1 dx_2$$
$$= \int g(x_1)\,p(x_1)\,dx_1 \cdot \int h(x_2)\,p(x_2)\,dx_2$$

:uncorrelated

$iii)$

$$YY^T = V^T X \left(V^T X\right)^T$$
$$= V^T XX^T V$$
$$= V^T V \Lambda V^T V$$
$$= I \Lambda I$$
$$= \Lambda.$$

Λ is a diagonal matrix composed of the eigenvalue of XX^T.

<u>Exercise 3: PCA – Derivation of the estimation of the PCs</u>

While in class we saw one way to compute the principal components of a distribution in one go, often it is practical to do it in an iterative manner. For this, one may iteratively minimize a functional. Following this idea,

a) show that maximizing the functional $J(w) = w^T C w$, under the constraint $\|w\| = 1$, where *C* is the covariance matrix and *w* a vector, finds an eigenvector of *C*.

b) Further, show that if *x* has a zero mean distribution, then the distribution of $y = w^T x$ the projection of *x* on the eigenvector has zero mean.

<u>Solution:</u>

a) $$\frac{\partial J(w) - \lambda\left(\|w\| - 1\right)}{\partial w} = 2Cw - 2\lambda w$$
$$\Rightarrow Cw = \lambda w$$

b) $\mathrm{E}\{y\} = \mathrm{E}\{w^T x\} = w^T \mathrm{E}\{x\} = 0$

Supplementary Exercises (to be done at home)

Exercise 1: Probabilities and Statistics

For what constant k is $f(x) = k\, e^{-x}$ a probability density function on [0,1]?

Solution:

We need to choose a k that makes requirements a) $f(x) >= 0$ and b) $\int f(x)dx=1$ of the definition of a density function true. Since $e^{-x} > 0$ for all x, all we need for (a) is to make sure that we choose $k \geq 0$. For (b), we calculate

$$\int_0^1 ke^{-x}\, dx$$
$$= [ke^{-x}]^1_0$$
$$= k(1 - 1/e) = k(e-1)/e..$$

Since this must equal 1, we get

$$k(e-1)/e = 1,$$

so

$$k = e/(e-1) \approx 1.582.$$

Therefore, the function

$f(x) = [\, e/(e-1)\,]e^{-x}$ is a probability density function on [0,1].

Exercise 2: Principal Component Analysis

Compute the principal components and eigenvalues of the following matrix:

$$A = \begin{bmatrix} 2 & -4 \\ -1 & -1 \end{bmatrix}$$

Solution:

$$\det\begin{bmatrix} 2-\lambda & -4 \\ -1 & -1-\lambda \end{bmatrix}$$
$$=(2-\lambda)(-1-\lambda)-(-4)(-1)$$
$$=\lambda^2-\lambda-6$$
$$=(\lambda-3)(\lambda+2)$$

Thus, $\lambda_1 = 3,\ \lambda_2 = -2$ are the eigenvalues of *A*. All eigenvectors corresponding to λ_1 = 3 are multiples of $[-4\ 1]^T$. An eigenvector corresponding to λ_2 = -2 is $[\ 1\ 1\ ]^T$

Exercise 3: Covariance, correlation

Show that $-1 \le corr(x, y) \le 1$. Recall that $\operatorname{var}\left(\frac{x}{\sigma_x} + \frac{y}{\sigma_y}\right) \ge 0$ and that $\operatorname{var}(x+y) = \sigma_x + \sigma_y - 2\operatorname{cov}(x, y)$ and that $corr(x, y) = \frac{\operatorname{cov}(x, y)}{\sigma_x \sigma_y}$.

Solution:

The variance of any quantity is nonnegative by definition, so

$$\operatorname{var}\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) \ge 0.$$

From a property of variances, the sum can be expanded

$$\operatorname{var}\left(\frac{X}{\sigma_X}\right) + \operatorname{var}\left(\frac{Y}{\sigma_Y}\right) + 2\operatorname{cov}\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right) \ge 0$$

$$\frac{1}{\sigma_X^2}\operatorname{var}(X) + \frac{1}{\sigma_Y^2}\operatorname{var}(Y) + \frac{2}{\sigma_X\sigma_Y}\operatorname{cov}(X, Y) \ge 0$$

$$1 + 1 + \frac{2}{\sigma_X\sigma_Y}\operatorname{cov}(X, Y) = 2 + \frac{2}{\sigma_X\sigma_Y}\operatorname{cov}(X, Y) \ge 0.$$

Therefore,

$$\operatorname{cor}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sigma_X\sigma_Y} \ge -1.$$

Similarly,

$$\operatorname{var}\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right) \ge 0$$

$$\operatorname{var}\left(\frac{X}{\sigma_X}\right) + \operatorname{var}\left(-\frac{Y}{\sigma_Y}\right) + 2\operatorname{cov}\left(\frac{X}{\sigma_X}, -\frac{Y}{\sigma_Y}\right) \ge 0$$

$$\frac{1}{\sigma_X^2}\operatorname{var}(X) + \frac{1}{\sigma_Y^2}\operatorname{var}(Y) - \frac{2}{\sigma_X\sigma_Y}\operatorname{cov}(X, Y) \ge 0$$

$$1 + 1 - \frac{2}{\sigma_X \sigma_Y} \operatorname{cov}(X, Y) = 2 - \frac{2}{\sigma_X \sigma_Y} \operatorname{cov}(X, Y) \geq 0.$$

Therefore,

$$\operatorname{cor}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sigma_X \sigma_Y} \leq 1,$$

so $-1 \leq \operatorname{cor}(X, Y) \leq 1$.