Chalmers tekniska högskola

TDA231 *Algorithms for Machine Learning & Interference*
Teacher: *Devdatt Dubhashi*

Homework 0

Victor Nilsson - *vicnilss@student.chalmers.se*

Bjarki Vilmarsson - *bjarkiv@student.chalmers.se*

23 januari 2017

1 Theoretical problems

Problem 1.1

The probability that a random person are cancer positive is $P(C_+) = 0.0001$ and thus the probability of being cancer negative is $P(C_-) = 0.9999$. The testing of the disease is not 100 % accurate and the probabiliy of actually getting a positive test given the patient being sick is $P(T_+|C_+) = 0.99$ and conversely the probability of getting negative result when being healthy is $P(T_-|C_-) = 0.99$. The definition of conditional probability is,

$$P(A|B) \triangleq \frac{P(A \cap B)}{P(B)}, \tag{1}$$

so the risk of actually being sick given a positive result is,

$$\begin{array}{rcl} P(C_+|T_+) & = & \dfrac{P(C_+ \cap T_+)}{P(T_+)} \\ & \Rightarrow & \\ P(C_+ \cap T_+) & = & P(C_+|T_+)P(T_+), \end{array} \tag{2}$$

and similarly we can write,

$$P(C_+ \cap T_+) = P(T_+|C_+)P(C_+), \tag{3}$$

and thus,

$$\begin{array}{rcl} P(T_+|C_+)P(C_+) & = & P(C_+|T_+)P(T_+) \\ & \Rightarrow & \\ P(C_+|T_+) & = & \dfrac{P(T_+|C_+)P(C_+)}{P(T_+)}. \end{array} \tag{4}$$

We can also use that

$$\begin{array}{rcl} P(T_+) & = & P(T_+|C_+)P(C_+) + P(T_+|C_-)P(C_-) \\ & = & P(T_+|C_+)P(C_+) + (1 - P(T_-|C_-))P(C_-) \end{array} \tag{5}$$

Inserting the given data we get,

$$\begin{array}{rcl} P(C_+|T_+) & = & \frac{0.99 \cdot 0.0001}{0.99 \cdot 0.0001 + 0.01 \cdot 0.9999} \\ & \approx & 0.0098 \end{array} \tag{6}$$

Answer: So the chance of actually being sick is only a mere 0.98 %.

Problem 1.2

The covariance of two random variables X and Y is defined as,

$$\begin{array}{rcl} Cov[X,Y] & \triangleq & E[(X - E[X])(Y - E[Y])] \\ & = & E[XY] - E[X]E[Y] \end{array} \tag{7}$$

.

For our problem we have that X is uniform on the interval $[-1,1]$ and $Y := X^2$, so

$$
\begin{aligned}
E[X] &= \int_{-1}^{1} x p_X(x) dx \\
\left\{ \begin{array}{rcl} \int_{-1}^{1} p(x)_X dx &=& 1 \\ \int_{-1}^{1} c dx &=& 1 \\ 2c &=& 1 \\ c &=& \frac{1}{2} \end{array} \right\} &= \frac{1}{2} \int_{-1}^{1} x dx \\
&= \frac{1}{4} \left[x^2 \right]_{-1}^{1} \\
&= 0.
\end{aligned}
\tag{8}
$$

Since $E[X]$ is zero we do not have to calculate the expectation of Y since it disappears anyway, though, we need to calculate the expectation of the product,

$$
\begin{aligned}
E[XY] &= \int_{-1}^{1} x p_{XY}(x) dx \\
\{ \; p_{XY}(x) \propto p_X^3(x) \; \} &\propto \int_{-1}^{1} x p_X^3(x) dx \\
&= 0.
\end{aligned}
\tag{9}
$$

If we insert this in Eq. (7) we get

$$
\begin{aligned}
Cov[X,Y] &= 0 - 0 \cdot E[Y] \\
&= 0
\end{aligned}
\tag{10}
$$

This means that the random variables X and Y are *uncorrelated* which does not imply *independence* since the variables obviously are.

2 Practical problems

Problem 2.1

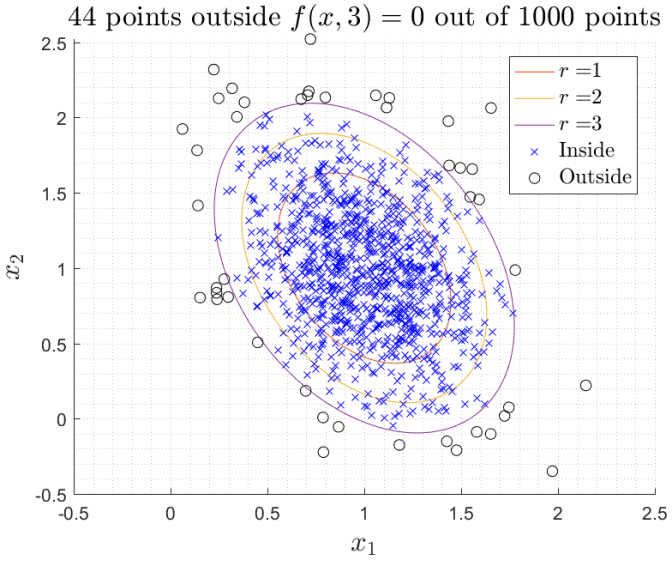
We want to study the distribution of a 2D multivariate normal distribution with mean $\mu = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and covariance $\Sigma = \begin{bmatrix} 0.1 & -0.05 \\ -0.05 & 0.2 \end{bmatrix}$.
The levels of the logarithm of the distribution follows,

$$f(x;r) := \frac{(x-\mu)^T \Sigma^{-1} (x-\mu)}{2} - r \tag{11}$$

where x is a random 2D vector and r is the level. Solving the equation

$$f(x;r) = 0, \tag{12}$$

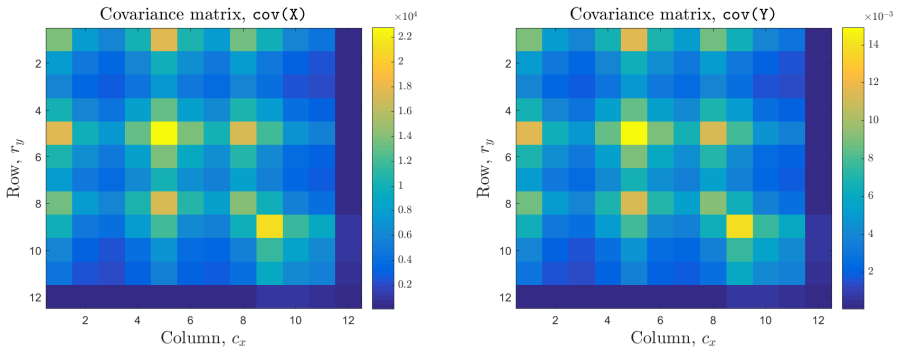
for different values of r yields level curves for this parameter which can be seen in Figures 1. In the same figure we can see 1000 randomly generated points using `mvnrnd(mu,Sigma,1000)` in MATLAB, we separate the points lying outside the level curve $f(x;r)$ from those that are inside.



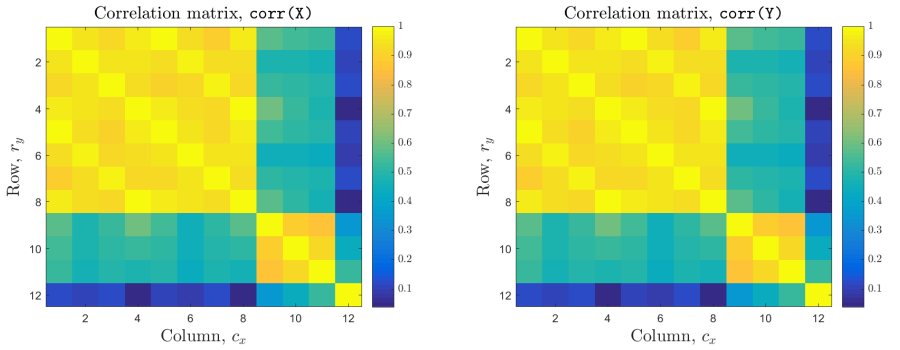
Figur 1: *Level set for the function given in Eq.* (11)*, together with randomly placed points following a 2D multivariate normal distribution.*

Problem 2.2

The correlation and covariance matrices for X are computed using the functions `cov()` and `corr()`. To scale each feature between 0 and 1 to make Y each value is divided by the highest number in the data set. The correlation and covariance matrices for both X and Y are plotted and shown in Figures 2. and 3.

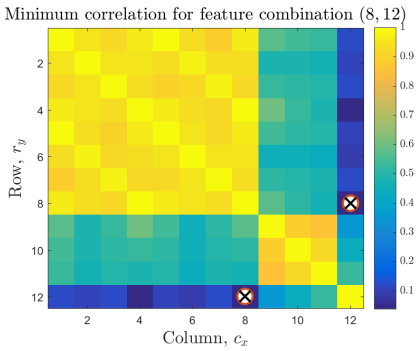


Figur 2: *The covariance of two random variables shows their joint variability. If great values in one variable corresponds to great values in the other and the same is true for small values then the covariance tend to be large. The covariance also contains information of the variance of the variables, therefore there is no surprise that the downscaled matrix Y contains much smaller elements.*



Figur 3: *The correlation of two variables is a measure of how much they correlate, ranging from* 1*, positive correlated, to* -1 *negative correlation. It is no surprise that the original matrix X has the same correlation matrix as the normalized matrix Y since all the numbers are scaled in the same manner.*

Both the covariance and correlation look identical for both X and Y when visualized. The correlation matrices are identical whereas the covariance matrices don't have the same numerical values but the ratio between values is the same. This was to be expected since Y is just a scaled version of X.
Lastly the pair with the minimum correlation are plotted. The pair is column 8 and 12. As seen in Figure 4 there are two clear clusters, one where the data points are very close together and the other where the data points are sparsely distributed.



Figur 4: *The pair of features with minimum correlation are* (8,12) *marked in the figure.*