ISTANBUL TECHNICAL UNIVERSITY
COMPUTER ENGINEERING DEPARTMENT

# BLG454E Learning From Data
# $1^{st}$ Assignment

Emir Erdem - 150120113

March 9, 2017

**Question 1.** Let T=1 denotes that a person is a terrorist (T=0 not terrorist)
and S=1 denotes that a scanned person is identified as terrorist (S=0 identified as upstanding citizen)

Probability that a person in the flight is a terrorist $P(T = 1) = 1\%$

Probability that a scanned terrorist is identified as terrorist $P(S = 1|T = 1) = 95\%$

Probability that a scanned upstanding citizen is identified as such
$P(S = 0|T = 0) = 95\%$ therefore $P(S = 1|T = 0) = 5\%$

It is asked that the probability that a person, who is identified as terrorist, is actually a terrorist $P(T = 1|S = 1) =?$

$$
\begin{aligned}
P(T=1|S=1) &= \frac{P(S=1|T=1)\times P(T=1)}{P(S=1)} \\
&= \frac{P(S=1|T=1)\times P(T=1)}{P(S=1|T=1)\times P(T=1)+P(S=1|T=0)\times P(T=0)} \\
&= \frac{95\%\times 1\%}{95\%\times 1\%+5\%\times 99\%} \\
&\cong 0.1610169 \\
&\cong 16.10169\%
\end{aligned}
$$

Consequently, with the probability of 16.101%, shifty looking man is a terrorist.

**Question 2.** The expected utility can be calculated as follows:

$$U(c(x^*)) = \sum_{c^{true}} U(c^{true}, c(x^*))\, p(c^{true}|x^*)$$

Optimal decision function $c(x^*)$ that maximizes the expected utility is

$$c(x^*) = \underset{c(x^*)}{\operatorname{argmax}}\, U(c(x^*))$$

Therefore, the expected utilities for each column can be found below.

$$
U(c(x^*)) = \begin{cases}
U(c^{true}=1, c(x^*)=5)\, p(c^{true}=1|x^*)+ & \\
U(c^{true}=2, c(x^*)=0)\, p(c^{true}=2|x^*)+ & \\
U(c^{true}=3, c(x^*)=-3)\, p(c^{true}=3|x^*), & \text{for } c(x^*)=1 \\
U(c^{true}=1, c(x^*)=3)\, p(c^{true}=1|x^*)+ & \\
U(c^{true}=2, c(x^*)=4)\, p(c^{true}=2|x^*)+ & \\
U(c^{true}=3, c(x^*)=0)\, p(c^{true}=3|x^*), & \text{for } c(x^*)=2 \\
U(c^{true}=1, c(x^*)=1)\, p(c^{true}=1|x^*)+ & \\
U(c^{true}=2, c(x^*)=-2)\, p(c^{true}=2|x^*)+ & \\
U(c^{true}=3, c(x^*)=10)\, p(c^{true}=3|x^*), & \text{for } c(x^*)=3
\end{cases}
$$

$$
U(c(x^*)) = \begin{cases}
5\times 0.7+0\times 0.2+(-3)\times 0.1, & \text{for } c(x^*)=1 \\
3\times 0.7+4\times 0.2+0\times 0.1, & \text{for } c(x^*)=2 \\
1\times 0.7+(-2)\times 0.2+10\times 0.1, & \text{for } c(x^*)=3
\end{cases}
$$

$$U(c(x^*)) = \begin{cases} 3.2, & \text{for } c(x^*) = 1 \\ 2.9, & \text{for } c(x^*) = 2 \\ 1.3, & \text{for } c(x^*) = 3 \end{cases}$$

Since the optimal function $c(x^*)$ has the highest value at the first column, the first column is the best decision to take.

**Question 3.** Gaussian probability density function

$$P(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu)^2\right], \qquad -\infty < x_i < \infty$$

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

Given N data samples (10 American female weights), the likelihood is

$$\mathcal{L}(\mu, \sigma) = P(x_{1:N}|\mu, \sigma) = \prod_{i=1}^{N} P(x_i|\mu, \sigma) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu)^2\right]$$

and the log-likelihood is

$$\log \mathcal{L}(\mu, \sigma) = \log \frac{1}{\sqrt{2\pi\sigma^2}} + \left(-\frac{1}{2\sigma^2}\right) \sum_{i=1}^{N} (x_i - \mu)^2$$

In order to find estimate $\hat{\mu}$, we need to take derivative of $\log \mathcal{L}(\mu, \sigma)$ with respect to $\mu$ and equate it to zero.

$$\frac{\partial}{\partial \mu} \log \mathcal{L}(\mu, \sigma) = \frac{\partial}{\partial \mu} \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} + \frac{\partial}{\partial \mu} \left(-\frac{1}{2\sigma^2}\right) \sum_{i=1}^{N} (x_i - \mu)^2 = 0$$

$$= \frac{1}{\sigma^2} \left[\sum_{i=1}^{N} (x_i - \mu)\right] = 0$$

$$= \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \mu = 0$$

Therefore, the estimate of $\mu$ and the mean value of the given weights is

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i = \frac{1}{10}(115 + 122 + 130 + 127 + 149 + 160 + 152 + 138 + 149 + 180) = 142.2$$

When the estimate is defined more generally, it is called estimator and written with capital X. Estimator of $\mu$ is $\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} X_i$

For the estimate $\hat{\sigma^2}$, we need to take derivative of $\log \mathcal{L}(\mu, \sigma)$ with respect to $\sigma^2$ and equate it to zero. In order to avoid possible mistakes, $\theta$ will be used instead of $\sigma^2$ throughout the equations.

$$\frac{\partial}{\partial \theta} \log \mathcal{L}(\mu, \theta) = \frac{\partial}{\partial \theta} \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\theta}} + \frac{\partial}{\partial \theta} \left( -\frac{1}{2\theta} \right) \sum_{i=1}^{N} (x_i - \mu)^2 = 0$$

$$= -\frac{N}{2\theta} + \frac{1}{2\theta^2} \left[ \sum_{i=1}^{N} (x_i - \mu)^2 \right] = 0$$

$$= -N\theta + \sum_{i=1}^{N} (x_i - \mu)^2 = 0$$

Thus, the estimate of $\theta$ and $\sigma^2$ is

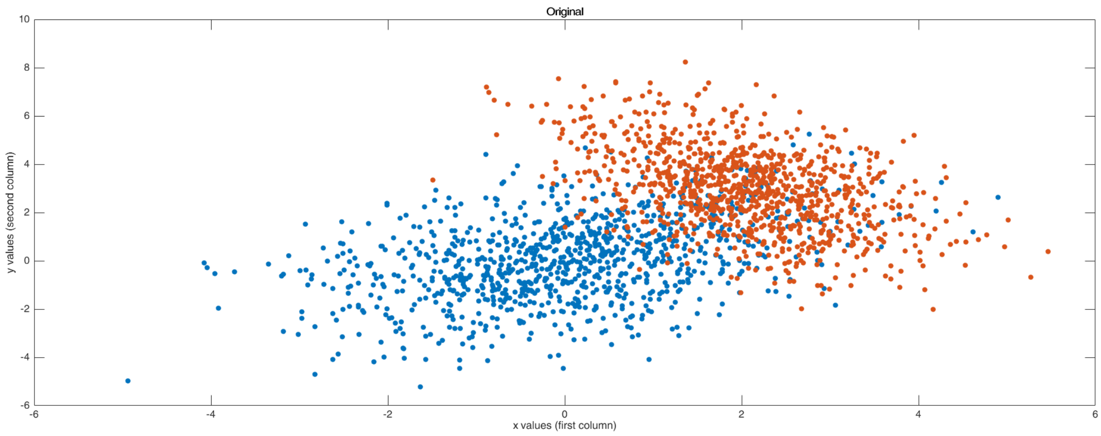
$$\hat{\theta} = \hat{\sigma^2} = \frac{\sum_{i=1}^{N} (x_i - \hat{\mu})^2}{N}$$

$$= \frac{(115 - 142.2)^2 + (122 - 142.2)^2 + \cdots + (180 - 142.2)^2}{10}$$

$$= 347.96$$

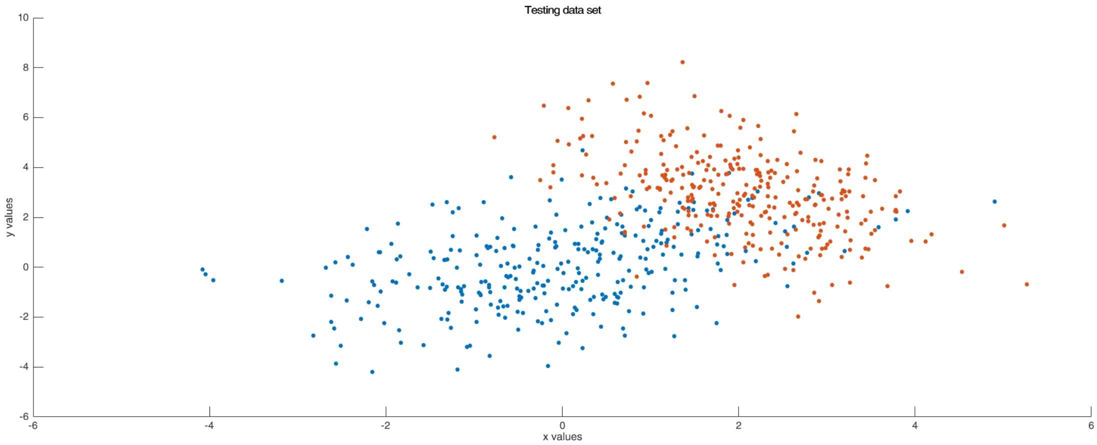
and the estimator of $\hat{\sigma^2} = \dfrac{\sum_{i=1}^{N} (X_i - \hat{\mu})^2}{N}$

**Question 4.** Original data is depicted below (blue = label 0, red = label 1).

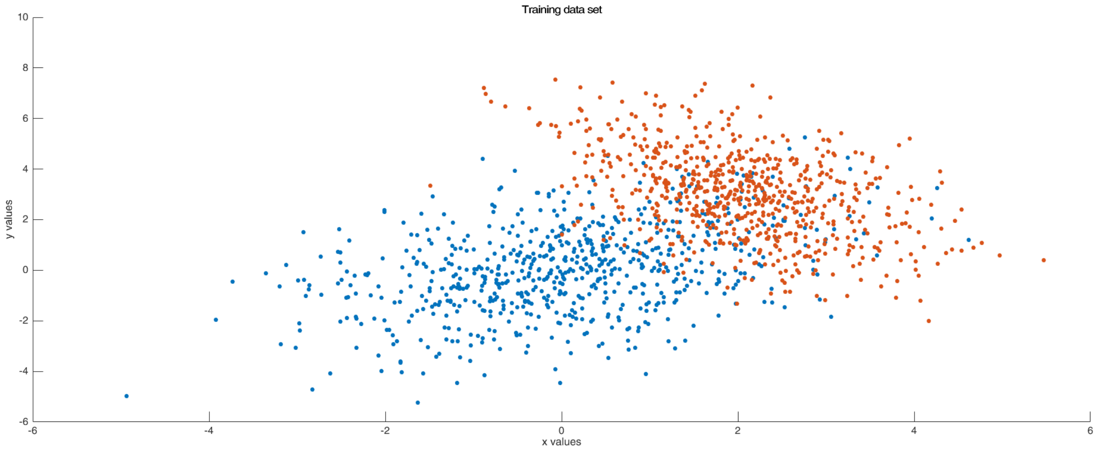


In order to examine this dataset, 30% of whole points are considered as test cases. From within total 2000 points, 600 test cases are picked out from the beginning of the dataset (300) and the end of the dataset (300).

Testing data set can be seen as follows.

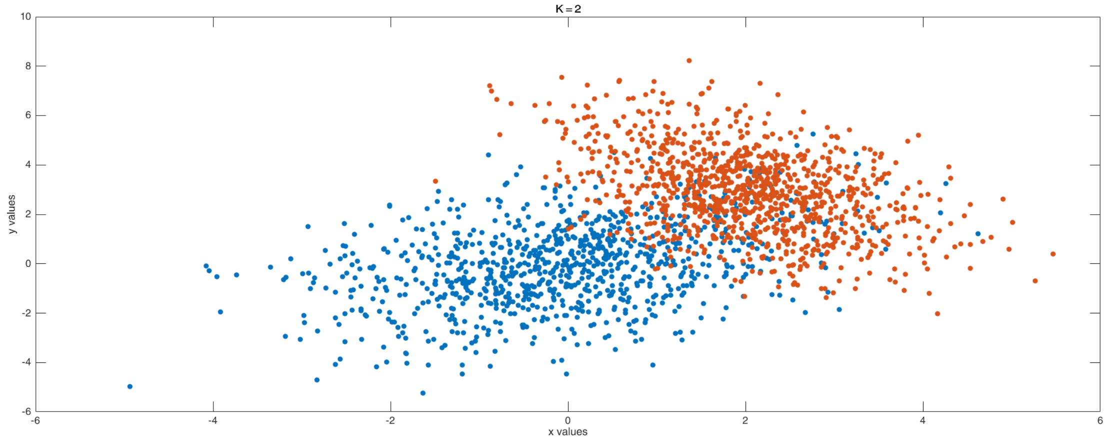



Training data set, which is the subtraction of testing data set from original data set, can be found below.

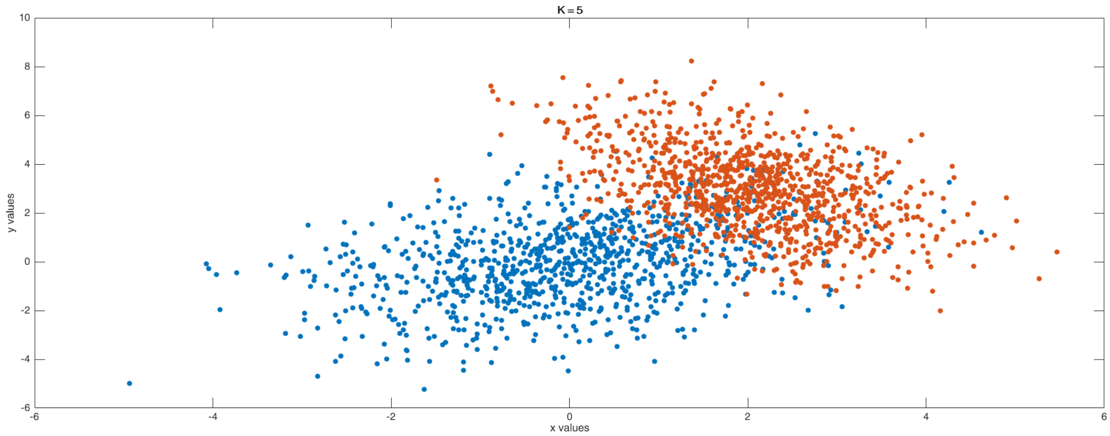



Every test point is classified using KNN classifier with 100 k-values (1-100).

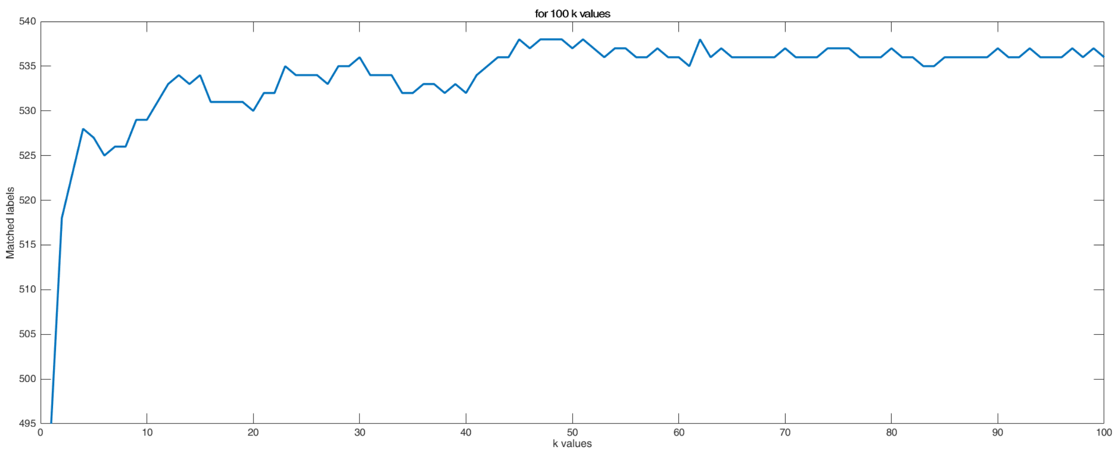
The figure below shows the classified data using KNN classification where K = 2.




The figure below shows the classified data using KNN classification where K = 5.




After all test points are classified by 100 k-values, their new labels are compared with the labels from original dataset. Matched labels are shown below for 100 k-values.



When all figures analyzed, it can be seen that when k value is increased, the matched labels are increased. Matched labels starts from 495 (k=1) and converges to approximately 537 (out of 600) after nearly 50 k-values. That means ~89.5% success.

Therefore, it can be deduced that increasing k value is useful up to a point.