Datasta Tietoon, Autumn 2007

SOLUTIONS TO EXERCISES 3

H3 / Problem 1.

See lectures slides, chapter 5. Given a data set $\mathbf{X} = (x(1), x(2), \ldots, x(n))$ and a model of a probability density function $p(x|\theta)$ with an unknown constant parameter vector θ, maximum likelihood method ("suurimman uskottavuuden menetelmä") estimates vector $\hat{\theta}$ which maximizes the likelihood function: $\hat{\theta}_{ML} = \max_\theta p(\mathbf{X}|\theta)$. In other words, find the values of θ which most probably have generated data $\mathbf{X}$.

Normally the data vectors $\mathbf{X}$ are considered independent so that likelihood function is a product of individual terms $p(\mathbf{X}|\theta) = p(x(1), x(2), \ldots, x(n)|\theta) = p(x(1)|\theta) \cdot p(x(2)|\theta) \cdot \ldots \cdot p(x(n)|\theta)$. Given a numerical data set $\mathbf{X}$, likelihood is function of only θ. Because the maximum of the likelihood $p(\mathbf{X}|\theta)$ and log-likelihood $\ln p(\mathbf{X}|\theta)$ is reached at the same value θ, log-likelihood function $L(\theta)$ is prefered for computational reasons. While $\ln(A \cdot B) = \ln A + \ln B$, we get $L(\theta) = \ln p(\mathbf{X}|\theta) = \ln \prod_j p(x(j)|\theta) = \sum_j \ln p(x(j)|\theta)$.

Remember also that $p(x, y|\theta)$ can be written with conditional probabilities $p(x, y|\theta) = p(x)p(y|x, \theta)$.

In this problem the model is $y(i) = \theta x(i) + \epsilon(i)$ which implies $\epsilon(i) = y(i) - \theta x(i)$. If there were no noise ϵ, θ could be computed from a single observation $\theta = y(1)/x(1)$. However, now the error ϵ is supposed to be zero-mean Gaussian noise with standard deviation σ: $\epsilon \sim N(0, \sigma)$, that is $E(\epsilon) = 0, Var(\epsilon) = \sigma^2$. This results to $E(y(i)|x(i), \theta) = \theta x(i) + E(\epsilon) = \theta x(i)$ and $Var(y(i)|x(i), \theta) = Var(\epsilon(i))$. Hence $(y(i)|x(i), \theta) \sim N(\theta x(i), \sigma)$ the density function is

$$p(y(i)|x(i), \theta) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(y(i)-\theta x(i))^2}{2\sigma^2}} \tag{1}$$

The task is to maximize $p(x, y|\theta) = p(x)p(y|x, \theta)$ w.r.t. θ. Assuming data vectors independent we get likelihood as $\prod_i p(x(i))p(y(i)|x(i), \theta)$. After taking logarithm the log-likelihood function is

$$L(\theta) = \text{const} + \sum_{i=1}^{n} \left(\ln \frac{1}{\sqrt{2\pi}\sigma} - \frac{(y(i) - \theta x(i))^2}{2\sigma^2}\right) \tag{2}$$

$$= \text{const} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y(i) - \theta x(i))^2 \tag{3}$$

Maximizing $L(\theta)$ is equal to minimizing its opposite number: $\min_\theta \frac{1}{2\sigma^2} \sum_{i=1}^n (y(i) - \theta x(i))^2 = \min_\theta \frac{1}{2\sigma^2} \sum_{i=1}^n (\epsilon(i))^2$. This equals to least squares estimation ("pienimmän neliösumman menetelmä") because of the certain properties of ϵ in this problem.

Minimum is fetched by setting the derivative w.r.t. θ to zero (the extreme point):

$$0 = \frac{\partial}{\partial \theta} \sum_{i=1}^{n} (y(i) - \theta x(i))^2 \tag{4}$$

$$= \sum_{i=1}^{n} \big(2(y(i) - \theta x(i))(-x(i))\big) \tag{5}$$

$$= -2 \sum_{i=1}^{n} y(i)x(i) + 2\theta \sum_{i=1}^{n} (x(i))^2 \tag{6}$$

$$\tag{7}$$

which gives finally the estimate

$$\hat{\theta}_{ML} = \frac{\sum_{i=1}^n x(i)y(i)}{\sum_{i=1}^n x(i)^2} \tag{8}$$

H3 / Problem 2.

See lectures slides, chapter 5, and Problem 1. Bayes rule is

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \tag{9}$$

Often only the maximum posterior estimate of θ (MAP) is computed. Taking logarithm gives $\ln p(\theta|x) = \ln p(x|\theta) + \ln p(\theta) - \ln p(x)$, and the derivative w.r.t. θ is set to zero: $\frac{\partial}{\partial \theta} \ln p(x|\theta) + \frac{\partial}{\partial \theta} \ln p(\theta) = 0$. Compared to ML-estimation (Problem 1), there is an extra term $\frac{\partial}{\partial \theta} \ln p(\theta)$.

In this problem we have also a data set $\mathbf{X}$ and now two variables θ and α to be estimated. The model is $y(i) = \alpha + \theta x(i) + \epsilon(i)$, where $\epsilon \sim N(0, \sigma)$ as in Problem 1. Now $E(y(i)|x(i), \alpha, \theta) = \alpha + \theta x(i)$, and $Var(y(i)|x(i), \alpha, \theta) = Var(\epsilon) = \sigma^2$. Thus $y(i) \sim N(\alpha + \theta x(i), \sigma)$ and the likelihood function is

$$p(y(i)|x(i), \alpha, \theta) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(y(i)-\alpha-\theta x(i))^2}{2\sigma^2}} \tag{10}$$

Parameters have also normal density functions ("prior densities")

$$\alpha \sim N(0, 0.1) \rightarrow p(\alpha) = \frac{1}{\sqrt{2\pi} \cdot 0.1} e^{-\frac{(\alpha-0)^2}{2 \cdot 0.1^2}} = \text{const} \cdot e^{-50\alpha^2} \tag{11}$$

$$\theta \sim N(1, 0.5) \rightarrow p(\theta) = \frac{1}{\sqrt{2\pi} \cdot 0.5} e^{-\frac{(\theta-1)^2}{2 \cdot 0.5^2}} = \text{const} \cdot e^{-2(\theta-1)^2} \tag{12}$$

In Bayes MAP-estimation the log posterior probability to be maximized is $\ln p(x, y|\alpha, \theta) + \ln p(\alpha) + \ln p(\theta)$, where the first term is the likelihood and the two latter terms prior densities:

$$\ln p(\alpha) = \text{const} - 50\alpha^2 \tag{13}$$

$$\ln p(\theta) = \text{const} - 2(\theta - 1)^2 \tag{14}$$

Hence, the task is

$$(\hat{\alpha}, \hat{\theta}) = \arg\max_{\alpha,\theta} \left\{ (-\frac{1}{2\sigma^2}) \sum_{i=1}^{n} \left[(y(i) - \alpha - \theta x(i))^2\right] - 50\alpha^2 - 2(\theta - 1)^2 \right\} \tag{15}$$

First, maximize w.r.t. α,

$$0 = \frac{\partial}{\partial \alpha} (-\frac{1}{2\sigma^2}) \sum_{i=1}^{n} \left[(y(i) - \alpha - \theta x(i))^2\right] - 50\alpha^2 - 2(\theta - 1)^2 \tag{16}$$

$$= (-\frac{1}{2\sigma^2}) \sum_i \left[2 \cdot (y(i) - \alpha - \theta x(i)) \cdot (-1)\right] - 100\alpha \tag{17}$$

$$= \sum_i y(i) - n\alpha - \theta \sum_i x(i) - 100\sigma^2\alpha \tag{18}$$

$$\hat{\alpha}_{MAP} = \frac{\sum_i y(i) - \theta \sum_i x(i)}{n + 100\sigma^2} \tag{19}$$

and similarly θ, using previous result of α,

$$0 = \frac{\partial}{\partial \theta} (-\frac{1}{2\sigma^2}) \sum_{i=1}^{n} \left[(y(i) - \alpha - \theta x(i))^2\right] - 50\alpha^2 - 2(\theta - 1)^2 \tag{20}$$

$$= (-\frac{1}{2\sigma^2}) \sum_i \left[2 \cdot (y(i) - \alpha - \theta x(i)) \cdot (-x(i))\right] - 4(\theta - 1) \tag{21}$$

$$= \sum_i \left[y(i)x(i) - \alpha x(i) - \theta x(i)^2\right] - 4\sigma^2(\theta - 1) \qquad | \quad \alpha \leftarrow \hat{\alpha}_{MAP} \tag{22}$$

$$= \sum_i y(i)x(i) - \left(\frac{\sum_i y(i) - \theta \sum_i x(i)}{n + 100\sigma^2}\right) \sum_i x(i) - \theta \sum_i x(i)^2 - 4\sigma^2\theta + 4\sigma^2 \tag{23}$$

$$\hat{\theta}_{MAP} = \frac{\sum_i y(i)x(i) - \frac{(\sum_i y(i))(\sum_i x(i))}{n+100\sigma^2} + 4\sigma^2}{\sum_i x(i)^2 - \frac{(\sum x(i))^2}{n+100\sigma^2} + 4\sigma^2} \tag{24}$$

Some interpretations of the results. If $\sigma^2 = 0$:

$$\begin{aligned}
\theta &= \frac{\sum_i y(i)x(i) - \frac{(\sum_i y(i))(\sum_i x(i))}{n}}{\sum_i x(i)^2 - \frac{(\sum x(i))^2}{n}} && (25)\\
&= (1/n) \cdot \frac{\sum_i y(i)x(i) - ((1/n)\cdot(\sum_i y(i)))((1/n)\cdot(\sum_i x(i)))}{(1/n)\cdot\sum_i x(i)^2 - ((1/n)\sum x(i))^2} && (26)\\
&= \frac{E(YX) - E(Y)E(X)}{E(X^2) - (E(X))^2} && (27)\\
&= \frac{Cov(X,Y)}{Var(X)} && (28)\\
\alpha &= (1/n)\sum_i y(i) - \theta(1/n)\sum_i x(i) && (29)\\
&= E(Y) - \theta E(X) && (30)
\end{aligned}$$

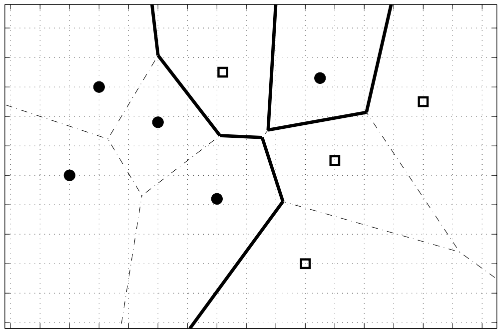
which are also the estimates of PNS method as well as by least squares.

If $\sigma^2 \to \infty$:

$$\begin{aligned}
\theta &\to 4/4 = 1 && (31)\\
\alpha &= \frac{\sum_i y(i) - \theta\sum_i x(i)}{n + 100\sigma^2} && (32)\\
&\to 0 && (33)
\end{aligned}$$

H3 / Problem 3.

1-NN border plotted with a thick line:



H3 / Problem 4.

Bayes rule $p(\omega|x) = \frac{p(x|\omega)p(\omega)}{p(x)}$.

Classification rule: when having observation x, choose class ω_1 if $p(\omega_1|x) > p(\omega_2|x) \Leftrightarrow \frac{p(x|\omega_1)p(\omega_1)}{p(x)} > \frac{p(x|\omega_2)p(\omega_2)}{p(x)}$ $\Leftrightarrow p(x|\omega_1)p(\omega_1) > p(x|\omega_2)p(\omega_2)$.

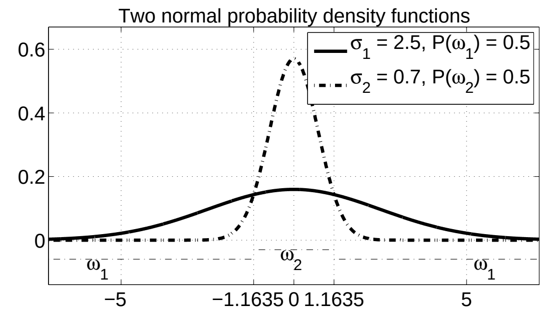
Now the both data follow the normal distribution $x|\omega_1 \sim N(0,\sigma_1)$ and $x|\omega_2 \sim N(0,\sigma_2)$. Assume that $\sigma_1^2 > \sigma_2^2$. See the density function curves in the figure below where $\sigma_1 = 2.5$ and $\sigma_2 = 0.7$ as an example. The density function of a normal distribution with mean μ and variance σ^2 is

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Now the rule is

$$\begin{aligned}
\frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{x^2}{2\sigma_1^2}} p(\omega_1) &> \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{x^2}{2\sigma_2^2}} p(\omega_2) && (34)\\
\frac{e^{-\frac{x^2}{2\sigma_1^2}}}{e^{-\frac{x^2}{2\sigma_2^2}}} &> \frac{\sigma_1}{\sigma_2}\frac{p(\omega_2)}{p(\omega_1)} \quad | \ \ln \text{ on both sides} && (35)\\
\Big(\frac{1}{2\sigma_2^2} - \frac{1}{2\sigma_1^2}\Big)x^2 &> \ln\Big(\frac{\sigma_1}{\sigma_2}\frac{p(\omega_2)}{p(\omega_1)}\Big) && (36)\\
x^2 &> \frac{2\ln(\frac{\sigma_1}{\sigma_2}\frac{p(\omega_2)}{p(\omega_1)})}{(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2})} && (37)
\end{aligned}$$

In the figure below the density functions and class borders when using sample values $\sigma_1 = 2.5$, $\sigma_2 = 0.7$, $P(\omega_1) = 0.5$, and $P(\omega_2) = 0.5$, yielding $x^2 > 1.3536$ and decision borders $|x| = 1.1635$. E.g., if we are given a data point $x = 2$, we choose the class ω_1.



H3 / Problem 5.

Probability of "1" is p and that of "0" is $1-p$. Then the probability of the vector "111010" is $p\cdot p\cdot p\cdot(1-p)\cdot p\cdot(1-p) = p^4(1-p)^2$.

a) There is a vector $\mathbf{x} = (x_1,\ldots,x_d)^T$, which has d elements, and the number of ones is N.

Now for the class ω_1, $p(x|\omega_1) = p^N(1-p)^{d-N}$ and correspondingly for the class ω_2, $p(x|\omega_2) = q^N(1-q)^{d-N}$

b) Suppose that $q < p$.

The classification rule: x belongs to ω_1, if $p(x|\omega_1)p(\omega_1) > p(x|\omega_2)p(\omega_2)$, or taking the logarithm $\ln p(x|\omega_1) + \ln p(\omega_1) > \ln p(x|\omega_2) + \ln p(\omega_2)$,

Substituting the density functions:

$$\begin{aligned}
\ln[p^N(1-p)^{d-N}] + \ln p(\omega_1) &> \ln[q^N(1-q)^{d-N}] + \ln p(\omega_2) && (38)\\
\ln p^N + \ln(1-p)^{d-N} + \ln p(\omega_1) &> \ln q^N + \ln(1-q)^{d-N} + \ln p(\omega_2) && (39)\\
N\ln p + (d-N)\ln(1-p) + \ln p(\omega_1) &> N\ln q + (d-N)\ln(1-q) + \ln p(\omega_2) && (40)\\
N[\ln p - \ln(1-p) - \ln q + \ln(1-q)] &> \ln p(\omega_2) - \ln p(\omega_1) + d\ln(1-q) - d\ln(1-p) && (41)\\
N\ln\Big(\frac{p(1-q)}{q(1-p)}\Big) &> \ln\Big(\frac{p(\omega_2)}{p(\omega_1)}\Big) + d\ln\Big(\frac{1-q}{1-p}\Big) \quad | \quad p(1-q)/(q(1-p)) > 1 \to \ln(.) > 0 && (42)\\
N &> \Big[\ln\frac{p(\omega_2)}{p(\omega_1)} + d\ln\Big(\frac{1-q}{1-p}\Big)\Big] \Big/ \Big[\ln\Big(\frac{p(1-q)}{q(1-p)}\Big)\Big] && (43)
\end{aligned}$$