Datasta Tietoon, Autumn 2007

SOLUTIONS TO EXERCISES 2

H2 / Problem 1.
a) See the figure below.
b)

$$E\{\mathbf{x}\} = \frac{1}{4}\sum \mathbf{x}(i) = \begin{bmatrix} 5 \\ 4 \end{bmatrix}$$

Thus the normalized data matrix is $\mathbf{X}_0 = \begin{bmatrix} -3 & 0 & 1 & 2 \\ -3 & -1 & 1 & 3 \end{bmatrix}$

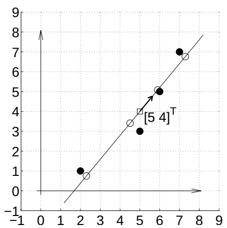
c) The covariance matrix is $\mathbf{C}_x = \frac{1}{4}\mathbf{X}_0\mathbf{X}_0^T = \frac{1}{4}\begin{bmatrix} 14 & 16 \\ 16 & 20 \end{bmatrix}$

The eigenvalues are computed from $\mathbf{C}_x\mathbf{u} = \lambda\mathbf{u}$, or by multiplying with 4, $\begin{bmatrix} 14 & 16 \\ 16 & 20 \end{bmatrix}\mathbf{u} = \mu\mathbf{u}$ where μ is 4 times λ. (It may be easier to solve the equation if the coefficients are integer numbers).

We have determinant $\begin{vmatrix} 14-\mu & 16 \\ 16 & 20-\mu \end{vmatrix} = 0$ which gives the characteristic equation $(14-\mu)(20-\mu) - 256 = 0$ or $\mu^2 - 34\mu + 24 = 0$. The roots are 33.28 and 0.72, hence the eigenvalues λ of the covariance matrix are these divided by 4.

The eigenvector corresponding to the larger eigenvalue can be computed by $\begin{bmatrix} 14 & 16 \\ 16 & 20 \end{bmatrix}\mathbf{u} = 33.28\mathbf{u}$ which (after some manipulation) gives $\mathbf{u} = [0.64\ 0.77]^T$.

The empty circles in the figure below are the projections onto 1D hyperplane (line), and $33.28/(33.28+0.72) \approx 97.9$ % of variance is explained.



H2 / Problem 2.
We can use the Lagrange optimization principle for a constrained maximization problem. The principle is saying that if we need to maximize $E\{(\mathbf{w}^T\mathbf{x})^2\}$ under the constraint $\mathbf{w}^T\mathbf{w} = 1$, we should find the zeroes of the gradient of

$$E\{(\mathbf{w}^T\mathbf{x})^2\} - \lambda(\mathbf{w}^T\mathbf{w} - 1)$$

where λ is the Lagrange constant.

We can write $E\{(\mathbf{w}^T\mathbf{x})^2\} = E\{(\mathbf{w}^T\mathbf{x})(\mathbf{x}^T\mathbf{w})\} = \mathbf{w}^T E\{\mathbf{x}\mathbf{x}^T\}\mathbf{w}$ because inner product is symmetrical and the E or expectation means computing the mean over the sample $\mathbf{x}(1), ..., \mathbf{x}(n)$, thus $\mathbf{w}$ can be taken out.

We need the following general result: if $\mathbf{A}$ is a symmetrical matrix, then the gradient of the quadratic form $\mathbf{w}^T\mathbf{A}\mathbf{w}$ equals $2\mathbf{A}\mathbf{w}$. It would be very easy to prove this by taking partial derivatives with respect to the elements of $\mathbf{w}$. This is a very useful formula to remember.

Now the gradient of the Lagrangian becomes:

$$2E\{\mathbf{x}\mathbf{x}^T\}\mathbf{w} - \lambda(2\mathbf{w}) = 0$$

or

$$E\{\mathbf{x}\mathbf{x}^T\}\mathbf{w} = \lambda\mathbf{w}$$

This is the eigenvalue - eigenvector equation for matrix $E\{\mathbf{x}\mathbf{x}^T\}$. But there are d eigenvalues and vectors: which one should be chosen?

Multiplying from the left by $\mathbf{w}^T$ and remembering that $\mathbf{w}^T\mathbf{w} = 1$ gives

$$\mathbf{w}^T E\{\mathbf{x}\mathbf{x}^T\}\mathbf{w} = \lambda$$

showing that λ should be chosen as the largest eigenvalue in order to maximize $\mathbf{w}^T E\{\mathbf{x}\mathbf{x}^T\}\mathbf{w} = E\{y^2\}$. This was to be shown.

H2 / Problem 3.
After convergence it must hold $\gamma[y\mathbf{x} - y^2\mathbf{w}] = 0$. Because $\gamma \neq 0$, it follows that either $y = 0$ or $\mathbf{x} - y\mathbf{w} = \mathbf{x} - (\mathbf{w}^T\mathbf{x})\mathbf{w} = 0$. In the former case, $\mathbf{w}$ becomes orthogonal to $\mathbf{x}$ because $\mathbf{w}^T\mathbf{x} = 0$.

In the latter case, $\mathbf{w}$ becomes aligned with $\mathbf{x}$. Denote $\mathbf{w} = \alpha\mathbf{x}$ and solve α: we have

$$\mathbf{x} - (\alpha\mathbf{x}^T\mathbf{x})\alpha\mathbf{x} = 0$$

which gives

$$\alpha = \frac{1}{\|\mathbf{x}\|}$$

Then finally

$$\mathbf{w} = \frac{\mathbf{x}}{\|\mathbf{x}\|}$$

or $\mathbf{w}$ tends to the unit vector in the orientation of $\mathbf{x}$.

(Actually, we can show that only the latter case is possible but it goes beyond elementary mathematics. For those of you who want to know: From the original update equation, by multiplying both sides with $\mathbf{x}^T$, we have

$$\mathbf{x}^T\mathbf{w} \leftarrow \mathbf{x}^T\mathbf{w} + \gamma[y\mathbf{x}^T\mathbf{x} - y^2\mathbf{x}^T\mathbf{w}]$$

or

$$\mathbf{x}^T\mathbf{w} \leftarrow \mathbf{x}^T\mathbf{w} + \gamma[(\mathbf{x}^T\mathbf{x})(\mathbf{x}^T\mathbf{w}) - (\mathbf{x}^T\mathbf{w})^3]$$

So, the change in the value of $\mathbf{x}^T\mathbf{w}$ at one step of the algorithm is equal to $\gamma[(\mathbf{x}^T\mathbf{x})(\mathbf{x}^T\mathbf{w}) - (\mathbf{x}^T\mathbf{w})^3]$. If $0 < \mathbf{x}^T\mathbf{w} < \|\mathbf{x}\|$, then the change is positive, meaning that $\mathbf{x}^T\mathbf{w}$ will *increase*. If on the other hand $\mathbf{x}^T\mathbf{w} > \|\mathbf{x}\|$, then the change is negative and $\mathbf{x}^T\mathbf{w}$ will *decrease*. In both cases, it will converge to $\|\mathbf{x}\|$, different from zero.)

H2 / Problem 4.
The likelihood function (supposing that given data samples $x(i)$ are independent; function of λ only)

$$p(\mathbf{x}|\lambda) = \prod_{i=1}^{n} p(x(i)|\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x(i)}$$

The log-likelihood:

$$L(\lambda) = \ln p(\mathbf{x}|\lambda) = \sum_{i=1}^{n}[\ln\lambda - \lambda x(i)] = n\ln\lambda - \lambda\sum_{i=1}^{n} x(i)$$

Putting the derivative with respect to λ to zero:

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\ln p(\mathbf{x}|\lambda) = n\frac{1}{\lambda} - \sum_{i=1}^{n} x(i) = 0$$

gives the solution

$$\frac{1}{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x(i).$$

Thus the ML (maximum likelihood) estimate for λ is the inverse of the mean value of the sample.

H2 / Problem 5.

The log-likelihood is

$$\ln p(\mathbf{x}|\mu,\sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}[x(j)-\mu]^2.$$

The log prior probability for μ is $\ln p(\mu) = \text{const} - \frac{1}{2}\mu^2$.

All the parts depending on μ in the Bayesian log posterior probability:

$$-\frac{1}{2\sigma^2}\sum_{j=1}^{n}[x(j)-\mu]^2 - \frac{1}{2}\mu^2$$

Putting the derivative w.r.t. μ to zero:

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}\mu}\left(-\frac{1}{2\sigma^2}\right)\sum_{j=1}^{n}[x(j)-\mu]^2 - \frac{1}{2}\mu^2 \\ &= -\frac{1}{2\sigma^2}\sum_{j=1}^{n}2[x(j)-\mu](-1) - \mu \end{aligned}$$

gives

$$\sum_{j=1}^{n}x(j) - n\mu - \sigma^2\mu = 0$$

which finally gives

$$\mu = \frac{1}{n+\sigma^2}\sum_{j=1}^{n}x(j).$$

The interpretation is as follows: if the variance σ^2 of the sample is very small, then μ is very close to the sample mean $\frac{1}{n}\sum_{j=1}^{n}x(j)$ because then the sample can be trusted.

On the other hand, if σ^2 is very large, then μ becomes close to zero which is the prior assumption. Then the sample cannot be trusted and the prior information dominates.