Tentamen: Introduction to Pattern Recognition (NB054B)

27 June, 2011, 10:30-13:30

Write your name and student number at the top of each sheet. On each page, indicate page number and total number of pages.

Please, write clearly! *Make sure to properly motivate all answers, and do not forget to include intermediate steps in your calculations: even a wrong answer may still gain you some points. You may refer to the Bishop book for relevant equations, etc.*

Assignment 1

A factory produces thingies. 8% of the thingies is of quality $x = 1$, 32% is of quality $x = 2$, and the remaining 60% is of quality $x = 3$. To assess the quality, two nondestructive tests have been developed that can be performed on a thingy. The first test, Z gives a numerical score z between 0 and 1. The conditional probability density of z depending on the quality x is

$$\begin{aligned} p(z|x=1) &= \alpha_1 z \\ p(z|x=2) &= \alpha_2(1-z) \\ p(z|x=3) &= \alpha_3(z+\exp(-z)) \end{aligned}$$

in which α_i, $i = 1, \ldots, 3$ are constants.

Question 1.1 *(a) Show that $\alpha_1 = 2$ and $\alpha_2 = 2$. (b) Compute α_3.*

ANSWER: Normalization constants from $\int p(z|x=1)dz = \int p(z|x=2)dz = 1$. Second as

$$\int p(z|x=1)dz = \alpha_2 \int_0^1 (1-z)dz = \alpha_2 \left[z - \frac{1}{2}z^2 \right]_0^1 = \alpha_2 \left[(1-0) - \frac{1}{2}(1-0) \right] = \frac{\alpha_2}{2} = 1$$

First likewise, so $\alpha_1 = 2$ and $\alpha_2 = 2$. (b) $\alpha_3 = (1.5 - 1/e)^{-1} \approx 0.8833$

Now there is another test Y. Its outcome is binary $y = 0$ or $y = 1$. The relation of y with the quality is deterministic:

$$\begin{aligned} p(y=1|x=1) &= 1 \\ p(y=1|x=2) &= 1 \\ p(y=1|x=3) &= 0 \end{aligned}$$

Question 1.2 *Compute $p(x = i|y = 1)$ for $i = 1, \ldots, 3$ using Bayes' rule*

ANSWER: Obviously, $p(x = 3|y = 1) = 0$, so you only have to compute the first two. Results in $p(x = 1|y = 1) = 0.2$ and $p(x = 2|y = 1) = 0.8$

Tests can be performed repeatedly on a thingy, and test outcomes are independent of each other if performed on the same thingy. Suppose we have performed both test Y and Z once.

Question 1.3 *Compute* $p(x = i|z, y = 1)$ *for* $i = 1, \ldots, 3$

ANSWER: Obviously, $p(x = 3|z, y = 1) = 0$. *So you only have to take the first two into account. Results in* $p(x = 1|z, y = 1) = \frac{z}{4-3z}$ *and* $p(x = 2|z, y = 1) = \frac{4-4z}{4-3z}$ *(see previous exam).*

We performed both tests Y and Z once, with result $y = 1$ and $z = 0.5$. We still want to have more certainty about the quality of the thingy. You have the choice to again perform Y, Z, or both.

Question 1.4 *Which of these tests, or combination of tests make sense and which not? For the test(s) you choose in this second round, do you expect to find a higher, lower or identical score than the first time? Motivate your answer.*

ANSWER: There is no sense in performing Y *again, since this will not further distinguish between class* $x = 1$ *and* $x = 2$. *The others are already excluded according to* $y = 1$. *Performing* Z *will provide new information, since outcome is independent of previous test results.*
After the results of the first round we have $p(x = 1|y = 1, z = 0.5) = \frac{z}{4-3z} = \frac{0.5}{2.5} = 0.2$ *and so* $p(x = 2|y = 1, z = 0.5) = \frac{4-4z}{4-3z} = \frac{2}{2.5} = 0.8$. *For* $x = 2$ *the chance on a score* $z < 0.5$ *is given by*

$$2\int_0^{0.5}(1-z)dz = 2\left[z - \frac{1}{2}z^2\right]_0^{0.5} = 2\left[(0.5-0) - \frac{1}{2}(0.25-0)\right] = 0.75$$

multiplied by the (updated, but unchanged) $p(x = 2) = 0.8$ *already shows that for a second test* Z *in at least 60% of the cases a smaller value of* z *will be obtained.*

Assignment 2

Consider a stochastic variable k that can have N outcomes $k = 1, \ldots, N$. We want to make a probability model $\vec{p} = (p_1, \ldots, p_N)$ for this variable, i.e. assign probabilities p_k to each of the possible outcomes k. We do not know $\vec{p}$, but suppose we do know that the expected value of a certain given function $f(k)$, denoted f_k for short, has the value F, i.e.

$$\langle f_k \rangle_{\vec{p}} \equiv \sum_{k=1}^{N} p_k f_k = F \tag{1}$$

Unfortunately, just knowing the expectation value $\langle f_k \rangle_{\vec{p}} = F$ is not sufficient to uniquely determine the probabilities $\vec{p}$.

Question 2.1 *Take* $N = 3$, $f_k = k$, *and* $F = 2$, *and show by example that there are at least two different distributions* $\vec{p}^{(1)}$ *and* $\vec{p}^{(2)}$ *that satisfy (1))*

ANSWER: The following distributions have $\langle f_k \rangle = 2$*: (1)* $\vec{p} = (0.5, 0, 0.5)$, *(2)* $p_2 = 1$ *rest zero, (3)* $p_i = \frac{1}{3}$, *etc.*

When confronted with a probability distribution in which only a few constraints are known, sometimes the *maximum entropy* (maxent) procedure is used. The entropy is defined as

$$H(\vec{p}) = -\sum_{k=1}^{N} p_k \ln p_k$$

Question 2.2 *Calculate the entropy for the two distributions you provided in question 1.*

ANSWER:

$$\begin{aligned} H(\vec{p} = (\frac{1}{2}, 0, \frac{1}{2})) &= -\ln 0.5 \approx 0.69 \\ H(\vec{p} = (0.0, 1, 0.0)) &= -\ln 1 = 0 \\ H(\vec{p} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})) &= -\ln \frac{1}{3} \approx 1.099 \end{aligned}$$

The idea is that one chooses the distribution that satisfies the constraints, but contains the least information otherwise. In our case, the probability vector is found by maximizing $H(\vec{p})$ under the constraints $\sum_{k=1}^{N} p_k f_k = F$ and $\sum_{k=1}^{N} p_k = 1$. For this we write down the Lagrangian with two Lagrange multipliers λ and μ,

$$L(\vec{p}, \lambda, \mu) = H(\vec{p}) + \lambda(\sum_{k=1}^{N} p_k - 1) + \mu(\sum_{k=1}^{N} p_k f_k - F) \tag{2}$$

To maximize L, we first maximize with respect to $\vec{p}$.

Question 2.3 *Set the gradient of the Lagrangian with respect to $\vec{p}$ equal to zero and verify that the probabilities in the maximum satisfy*

$$-\ln(p_k) - 1 + \lambda + \mu f_k = 0 \tag{3}$$

ANSWER: Take the partial derivative to p_k, this yields the equation. (On the exam, this should be elaborated, of course).

Now λ can be elimated using the normalisation constraint

Question 2.4 *Show, starting from (3) and eliminating λ that the maxent probabilities are of the form*

$$p_k = \frac{\exp(\mu f_k)}{\sum_{j=1}^{N} \exp(\mu f_j)} \tag{4}$$

ANSWER:

$$p_k = \exp(-1 + \lambda) \exp(\mu f_k)$$

Since $\sum p_k = 1$ *(normalisation)* $\exp(-1+\lambda) \sum_{j=1}^{N} \exp(\mu f_j) = 1$, *so* $\exp(-1+\lambda) = 1/(\sum_{j=1}^{N} \exp(\mu f_j))$.

Question 2.5 *(a) Argue why, for the given situation (with $N = 3$, $f_k = k$, and $F = 2$), the distribution $\vec{p} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ is the 'obvious' maxent solution; (b) Calculate the maximum entropy solution in case the expectation value is decreased to $F = 1.5$.*

ANSWER: (a) All outcomes equally probable gives the maximum possible entropy without any additional constraints (Bishop, p.52). As this solution also happens to satisfy the constraint $F = 2$ it is obviously the optimal solution for that case as well. (b) We still have to eliminate μ from (4), *which can be accomplished using the constraint $F = 1.8$ in* (1). *With $f_k = k$ this results in*

$$\begin{aligned} F &= \langle f_k \rangle_{\vec{p}} = \sum_{k=1}^{N} p_k k \\ &= \big(\exp(\mu) + 2\exp(2\mu) + 3\exp(3\mu)\big) / \sum_{j=1}^{3} \exp(j\mu) \\ &= \big(\alpha + 2\alpha^2 + 3\alpha^3\big) / \big(\alpha + \alpha^2 + \alpha^3\big) \end{aligned}$$

where we substituted $\alpha = \exp(\mu)$. *Dividing by* α *and rearranging then gives the quadratic equation*

$$(3 - F)\alpha^2 + (2 - F)\alpha + (1 - F) = 0$$

which can be solved for $F = 1.5$ *to give* $\alpha = \frac{1}{6}(\sqrt{13} - 1)$ *or* $\mu \approx -0.8341$. *Substituting this in* (4) *then gives* $\vec{p} \approx (0.6162, 0.2676, 0.1162)$. *It is easily verfified that under this distribution the expectation value is indeed* $\langle k \rangle = 1.5$.

THERE ARE TWO MORE QUESTIONS ON PAGE 3 AND 4

Assignment 3

Consider a Gaussian distribution $p(x, y, z)$ with mean

$$\boldsymbol{\mu} = (1, 3, 5)^T \tag{5}$$

and covariance

$$\boldsymbol{\Sigma} = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 4 \end{pmatrix} \tag{6}$$

Since $p(x, y, z)$ is a Gaussian distribution, we know that $p(x, z)$ should also be a Gaussian with a certain mean vector and covariance matrix.

Question 3.1 *What are the mean and covariance of the distribution $p(x, z)$? Sketch a few contours of constant probability.*

ANSWER: $\mu = (1, 5)$

$$\Sigma = \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix}$$

Sketch is similar to fig.2.8b with the axes reversed. (Should be drawn on the exam, of course)

Next we look at the probability distribution of $p(x|z)$.

Question 3.2 *(a) Is this a marginal, conditional or joint distribution? Explain. (b) Give an expression for the distribution of $p(x|z)$.*

ANSWER: The covariance is diagonal, so x and z are independent. So $p(x|z)$ is $p(x)$, so $\mu = 1$ and variance is $\sigma^2 = 2$. Direct computation using (2.96) is of course also OK.

We now consider a *different* Gaussian distribution $p(u, v, w) = p(u)p(v|u)p(w|u)$, defined in terms of conditional distributions as

$$p(u) = \mathcal{N}(u|\mu_0, \sigma^2) \tag{7}$$
$$p(v|u) = \mathcal{N}(v|c{\cdot}u, s^2) \tag{8}$$
$$p(w|u) = \mathcal{N}(w|d{\cdot}u, t^2) \tag{9}$$

with μ_0, σ^2, c, s^2, d and t^2 constant model parameters. The conditional distribution $p(u|v)$ is a Gaussian with a certain mean $\mu_{u|v}$ and variance $\sigma^2_{u|v}$. The question is: what are they?

Question 3.3 *Show that mean $\mu_{u|v}$ and variance $\sigma^2_{u|v}$ of the distribution $p(u|v)$ are given by*

$$\mu_{u|v} = \frac{\dfrac{\mu_0}{\sigma^2} + \dfrac{cv}{s^2}}{\dfrac{1}{\sigma^2} + \dfrac{c^2}{s^2}} \tag{10}$$

$$\frac{1}{\sigma^2_{u|v}} = \frac{1}{\sigma^2} + \frac{c^2}{s^2} \tag{11}$$

ANSWER: Fill in 2.116 and 2.117 (should be elaborated, of course)

The conditional distribution $p(u|v, w)$ is also a Gaussian.

Question 3.4 *Give an expression for the distribution $p(u|v,w)$ in terms of the model parameters.*

ANSWER: Fill in 2.116 and 2.117 gives $p(u|v,w) = \mathcal{N}(u|\mu_{u|vw}, \sigma^2_{u|vw})$, with

$$\mu_{u|vw} = \frac{\frac{\mu_0}{\sigma^2} + \frac{cv}{s^2} + \frac{dw}{t^2}}{\frac{1}{\sigma^2} + \frac{v^2}{s^2} + \frac{w^2}{t^2} +}$$

$$\frac{1}{\sigma^2_{u|vw}} = \frac{1}{\sigma^2} + \frac{c^2}{s^2} + \frac{d^2}{t^2}$$

THERE IS ONE MORE QUESTION ON PAGE 4

Assignment 4

In the early 80's Hinton and Sejnowski introduced the Boltzmann machine as a class of recurrent neural networks. Boltzmann machines can be understood as probability distributions of binary variables $\vec{s} = (s_1, \ldots, s_N)$, where $s_i = \pm 1$. The Boltzmann machine is parameterized by the matrix W, in which an element w_{ij} represents the connection strength between variables s_i and s_j. The distribution modeled by the Boltzmann machine is defined by

$$P(\vec{s}|W) = \frac{1}{Z(W)} \exp(\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij} s_i s_j) \tag{12}$$

with $Z(W)$ the normalisation constant,

$$Z(W) = \sum_{s_1=\pm 1} \ldots \sum_{s_N=\pm 1} \exp(\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij} s_i s_j) \tag{13}$$

Question 4.1 *Explain how a Boltzmann machine models that certain variables are likely to be both 'on' or 'off' $(+1/-1)$ together, have opposite values or have no relation at all.*

ANSWER: Note that $s_i = s_j \Leftrightarrow s_i s_j = 1$ and $s_i \neq s_j \Leftrightarrow s_i s_j = -1$ We rewrite $P(\vec{s}|W) \propto \prod_{ij} \exp(w_{ij} s_i s_j)$ and consider one of its factors $\exp(w_{ij} s_i s_j)$. It can basically take two values, namely $\exp(w_{ij})$ if $s_i = s_j$, and $\exp(-w_{ij})$ if $s_i \neq s_j$ Now if $w_{ij} > 0$, then $\exp(w_{ij}) > \exp(-w_{ij})$, so this favors values with $s_i = s_j$ and if $w_{ij} < 0$, then $\exp(w_{ij}) < \exp(-w_{ij})$, so this favors values with $s_i \neq s_j$.

Suppose that we have an i.i.d[1] data set $D = \{\vec{s}^1, \ldots, \vec{s}^M\}$. Note that each observation is a vector $\vec{s}^m = (s_1^m, \ldots, s_N^m)$. The goal is to learn the parameters in the connectivity matrix W from the data. For that, Hinton & Sejnowski introduced their famous Boltzmann machine learning rule,

$$\Delta w_{ij} = \eta \left[\langle s_i s_j \rangle_{\text{Data}} - \langle s_i s_j \rangle_{P(\vec{s}|W)} \right] \tag{14}$$

So learning is proportional to the difference of two terms. The first term is the empirical correlation between s_i and s_j in the data set:

$$\langle s_i s_j \rangle_{\text{Data}} \equiv \frac{1}{M} \sum_{m=1}^{M} s_i^m s_j^m \tag{15}$$

The second term is the correlation s_i and s_j under the current model $P(\vec{s}|W)$:

$$\langle s_i s_j \rangle_{\text{W}} \equiv \sum_{s_1=\pm 1} \ldots \sum_{s_N=\pm 1} s_i s_j P(\vec{s}|W) \tag{16}$$

We will derive the Boltzmann machine learning rule as gradient ascent on the loglikelihood.

Question 4.2 *Show that the log-likelihood of $L(W)$ for the data set D is given by*

$$L(W) = M(\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij} \langle s_i s_j \rangle_{Data} - \ln Z(W)) \tag{17}$$

with $\langle s_i s_j \rangle_{Data}$ as defined in (15) and $Z(W)$ as in (13).

ANSWER: For the likelihood we have

$$p(D|W) = \prod_{m=1}^{M} P(\vec{s}^m|W)$$

[1] independent and identically distributed

Taking the log, converting products to sums and collecting terms this gives

$$\begin{aligned}
\ln p(D|W) &= \ln \prod_{m=1}^{M} \left(\exp(\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij} s_i^m s_j^m) Z(W) \right) \\
&= \sum_{m=1}^{M} \Big(\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij} s_i^m s_j^m - \ln(Z(W)) \Big) \\
&= M \left(\frac{1}{M} \sum_{m=1}^{M} \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij} s_i^m s_j^m - \ln(Z(W)) \right) \\
&= M \left(\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij} \Big(\frac{1}{M} \sum_{m=1}^{M} s_i^m s_j^m \Big) - \ln(Z(W)) \right) \\
&= M(\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij} \langle s_i s_j \rangle_{Data} - \ln Z(W))
\end{aligned}$$

Now we have to take the gradient of $\frac{1}{M} L(W)$. Let us first consider the term with $\ln(Z(W))$,

Question 4.3 *Show that*

$$\frac{\partial}{\partial w_{ij}} \ln Z(W) = \langle s_i s_j \rangle_W$$

with $\langle s_i s_j \rangle_W$ defined as in (16).

ANSWER: The trick is $\frac{\partial}{\partial w_{ij}} \ln Z(W) = \frac{1}{Z(W)} \frac{\partial}{\partial w_{ij}} Z(W)$. (you should elaborate this further on the exam)

Question 4.4 *Combine these results to show that the gradient ascent on the log-likelihood (17)*

$$\Delta w_{ij} = \frac{\eta}{M} \frac{\partial}{\partial w_{ij}} L(W)$$

leads to the Boltzmann machine learning rule (14).

ANSWER: Collect all the terms... (you should write this out on the exam)