GI12/4C59 – Homework #4 (Due 12am, December 13, 2005)

Aim: To get familiarity with the basic concepts of clustering, maximum likelihood reasoning, model selection, and Gaussian channels. Presentation, clarity, and synthesis of exposition will be taken into account in the assessment of these exercises. This document is available from

http://www.cs.ucl.ac.uk/staff/Rob.Smith/Internal/information_theory.htm

1) **[30 pts]**
   a) A photon counter is pointed at a remote star for one minute, in order to infer the brightness, i.e., the rate of photons arriving at the counter per minute, *λ*. Assuming the number of photons collected *r* has a Poisson distribution with mean *λ,*

$$P(r \mid \lambda) = \exp(-\lambda)\frac{\lambda^r}{r!}$$

what is the maximum likelihood estimate for *λ,* given *r*=9?

   b) Same situation, but now we assume that the counter detects not only photons from the star but also ‘background’ photons. The background rate of photons is known to be *b*=13 photons per minute. We assume the number of photons collected, *r*, has a Poisson distribution with mean *λ*+*b*. Given *r*=9, what is the maximum likelihood estimate for *λ*?

2) **[30 pts]** Random variables *x* come independently from a probability distribution *P(x)*. According to model M_0, *P(x)* is a uniform distribution:

$$P(x \mid m, M_0) = \frac{1}{2} \qquad x \in \{-1,1\}$$

According to model M_1, *P(x)* is a non-uniform distribution with an unknown parameter $m \in \{-1,1\}$:

$$P(x \mid m, M_1) = \frac{1}{2}(1+mx) \qquad x \in \{-1,1\}$$

Given data *D={0.3,0.5,0.7,0.8,0.9}*, what is the evidence for M_0 and M_1?

3) **[20 pts]** You need to buy a modem for a rural house. The come in various kps (1000 bit/sec) ratings, and cost more for higher rates. You look up stats for the noisy telephone lines in the area, and find out they have bandwidth of 3kHz and signal-to-noise ratio of 30 dB. What’s the maximum kps modem worth buying?

4) **[20 pts]** Show that as the parameter *β* goes to infinity, the soft K means algorithm becomes the (hard) K means algorithm, except for the way in which cluster means with no assigned points behave. Describe that difference.

Model Answers:

1) **[30 pts]**

a)

$$\frac{\partial P(r\,|\,\lambda)}{\partial \lambda}=-\exp(-\lambda)\frac{\lambda^{r}}{r!}+r\exp(-\lambda)\frac{\lambda^{r-1}}{r!}=0$$

(note that we could have examined log likelihood, with the same results)

Ignoring trivial solutions, this derivative is zero when:

$$r-\lambda=0$$

Therefore, the maximum likelihood is given when $\lambda=9$.

b)

Same calculation as above, yielding:

$$r-(\lambda+b)=0$$

However, this yields the solution that $\lambda=-4$. This requires an interpretation of what “maximum likelihood” in this situation actually means. Are we reading a star with negative brightness (some sort of black hole)!

What is the valid range of values for λ? The original Poisson distribution is defined for positive numbers of occurrences (and is normalized as a probability distribution in that range). Therefore, the appropriate interpretation is that the maximum likelihood occurs in the range $\lambda \geq 0$. A little inspection shows that the maximum likelihood value is therefore λ=0.

2) **[30 pts]**

The evidence is simply the probability of the data given the model.

For

$$P(x\,|\,m,M_0)=\frac{1}{2} \qquad x\in\{-1,1\}$$

$$D=\{0.3,0.5,0.7,0.8,0.9\}$$

$$P(D\,|\,M_0)=P(0.3\,|\,M_0)P(0.5\,|\,M_0)P(0.7\,|\,M_0)P(0.8\,|\,M_0)P(0.9\,|\,M_0)$$

Let’s assume some accuracy of plus or minus *ε/2* in the measurements of *D.* This gives:

$$P(D\,|\,M_0)=\left(\frac{\varepsilon}{2}\right)^5=0.03125\varepsilon^5$$

For the other model:

$$P(x\,|\,m,M_1)=\frac{1}{2}(1+mx) \qquad x\in\{-1,1\}$$

In this case, we need to fit the model to the data, by determining the maximum likelihood value of *m*.

$$P(D\,|\,m,M_1)=\frac{1}{32}(1+0.3m)(1+0.5m)(1+0.7m)(1+0.8m)(1+0.9m)$$

$$=\frac{1}{32}\left(1+3.2m+3.98m^2+2.392m^3+0.6897m^4+0.0756m^5\right)$$

$$\frac{\partial P(D\,|\,m,M_1)}{\partial m}=\frac{1}{32}\left(3.2+7.96m+7.176m^2+2.7588m^3+0.378m^4\right)$$

$$=0.1+0.24875m+0.22425m^2+0.0862125m^3+0.0118125m^4$$

Taking the roots of this quartic (tools are available online, or you could use Mathematica) shows that all minima are outside $m\in\{-1,1\}$. Therefore, the maximum likelihood best fit of this model is at *m*=1, by inspection.

To determine the evidence for this model, integrating yields:

$$P\left(x_1 \mid m, M_1\right) = \int_{x_1-\frac{\varepsilon}{2}}^{x_1+\frac{\varepsilon}{2}} \left(\frac{1}{2}(1+x)\,dx\right)$$

$$= \frac{1}{2}\left(x+\frac{1}{2}x^2\right)\Bigg|_{x_1-\frac{\varepsilon}{2}}^{x_1+\frac{\varepsilon}{2}} = \left(\frac{1}{2}\left(x_1+\frac{\varepsilon}{2}+\frac{1}{2}\left(x_1+\frac{\varepsilon}{2}\right)^2\right)\right)-\left(\frac{1}{2}\left(x_1-\frac{\varepsilon}{2}+\frac{1}{2}\left(x_1-\frac{\varepsilon}{2}\right)^2\right)\right)$$

$$= \left(\frac{1}{2}\left(x_1+\frac{\varepsilon}{2}+\frac{1}{2}x_1^2+x_1\varepsilon+\frac{\varepsilon^2}{4}\right)\right)-\left(\frac{1}{2}\left(x_1+\frac{\varepsilon}{2}+\frac{1}{2}x_1^2-x_1\varepsilon+\frac{\varepsilon^2}{4}\right)\right)$$

$$= 2x_1\varepsilon$$

$$P\left(D \mid M_0\right) = 2.4192\varepsilon^5$$

Therefore, the second model has the greater evidence.

3) **[20 pts]**

$$C = b\log_2\left(1+\frac{v}{\sigma^2}\right)$$

Where *b* is bandwidth. To convert *dB* to a signal-to-noise ratio:

$$30dB = 10\log_{10}\left(\frac{v}{\sigma^2}\right)$$

$$\left(\frac{v}{\sigma^2}\right) = 1000$$

$$C = 3000\log_2\left(1+1000\right) = 29.902\text{kps}$$

4) **[20 pts]**

Consider the responsibility factor in soft K-means:

$$r_k^{(n)} = \frac{\exp\left(-\beta d\left(\mathbf{m}^{(k)},\mathbf{x}^{(n)}\right)\right)}{\sum_{k'}\exp\left(-\beta d\left(\mathbf{m}^{(k')},\mathbf{x}^{(n)}\right)\right)}$$

$$= \frac{1}{\sum_{k'}\exp\left(-\beta\left(d\left(\mathbf{m}^{(k')},\mathbf{x}^{(n)}\right)-d\left(\mathbf{m}^{(k)},\mathbf{x}^{(n)}\right)\right)\right)}$$

Let’s first assume that cluster *k* is the closest to point *n*. In this case, there is one term in the denominator that is exp(0)=1, and the remaining terms all have negative exponents. As *β* goes to infinity, each of these terms goes to $1/\infty = 0$. Thus, this yields *r*=1/1=1.

Next, if we assume that cluster *k* is not the closest to point *n*. In this case, as *β* goes to infinity, there are a number of terms in the denominator that go to zero, one that goes to one, and at least one that goes to infinity (the term associated with the closest cluster to the current point) . Thus, this yields $r = 1/\left(1+\infty\right) = 1/\infty = 0$.

Thus, soft K-means reduces to hard K-means in the limit as β goes to infinity, in terms of updates, with the exception of clusters that are not closest to any points. In hard K-means, these clusters receive no update, and in soft K-means they receive a weighted update.