

## Statistical Methods for Artificial Intelligence, Autumn 2006

### Problem set 1, Due Wednesday Oct. 4

Problem 1.

a) Suppose that there is a popular 10 megapixel digital camera where a pixel is represented by 16 bits. Consider the probability distribution over images to be taken by this model of camera over the next year (we call these “natural” images). Give an upper bound on the entropy of this distribution. Ignore the question of whether the distribution of natural images is really well defined.

b) Suppose that we use rendering software to construct images of solid models where a model consists of a set of objects each at a certain configuration and under certain lighting conditions and from a certain camera position. Suppose that the models are generated by kids using modeling software on the web where each model must fit in a single ten kilobyte message. Consider the probability distribution over the images rendered in this process. Give an upper bound on the entropy of this distribution. (Again ignore the question of whether the distribution is well defined).

c) Consider climate simulation software that samples future weather patterns by using a random number generator to add noise to the process of weather formation. If the program always starts with the same current state but uses a 32 bit random number seed (selected uniformly from all such seeds) what is the entropy of the probability distribution over future weather patterns produced by this program assuming that each different random seed produces a different weather pattern.

Problem 2.

a) A function  $f$  is convex over an interval  $(a, b)$  if for every  $x_1, x_2 \in (a, b)$  and  $0 \leq \lambda \leq 1$ ,

$$f(\lambda x_2 + (1 - \lambda)x_1) \leq \lambda f(x_2) + (1 - \lambda)f(x_1)$$

A function is convex if it is “upward curving”. A sufficient condition for convexity is that the second derivative exists and is non-negative (as in an

upward curving parabola). As  $\lambda$  moves from 0 to 1 the right hand side above gives a straight line (as a function of  $\lambda$ ) from  $f(x_1)$  to  $f(x_2)$ . On the other hand the left hand side is the value of function  $f$  on a corresponding point between  $x_1$  and  $x_2$ . A function is convex (upward curving) if, for any  $x_1$  and  $x_2$ , the function values between  $x_1$  and  $x_2$  are never above the straight line from  $f(x_1)$  to  $f(x_2)$ . Again consider an upward-curving parabola.

Let  $x$  be a real valued random variable taking values in a discrete set  $\{v_1, v_2, \dots, v_m\}$  and let  $p_i$  be the probability that  $x = v_i$ . show that:

$$f\left(\sum_{i=1}^m p_i v_i\right) = f(E_{x \sim p}[x]) \leq E_{x \sim p}[f(x)] = \sum_{i=1}^m p_i f(v_i)$$

This very useful relation is known as *Jensen's inequality*. Although we prove it here only for discrete distributions, it holds for continuous distributions as well where  $E_{x \sim p}[f(x)]$  is defined by an integral rather than a sum. Hint: Since we are only proving the discrete case, you can use induction on  $m$ .

b) Define the cross-entropy  $H(p, q)$  as follows.

$$H(p, q) = \sum_x p(x) \log \frac{1}{q(x)}$$

The cross entropy  $H(p, q)$  is the number of bits per message when we draw from distribution  $p$  but use a code for that is optimal for  $q$ . Define the *relative entropy* (also called KL (Kullback-Leibler) divergence) of a distribution  $q$  with respect to  $p$  as:

$$KL(p, q) = H(p, q) - H(p) = \sum_x p(x) \left( \log \frac{1}{q(x)} - \log \frac{1}{p(x)} \right)$$

$KL(p, q)$  is the extra coding cost when using the code for  $q$  rather than  $p$  when drawing from  $p$ . Use the fact that  $-\log(z)$  is a convex function of  $z$ , let  $z$  be the random quantity  $q(x)/p(x)$  (for random variable  $x$ ), and apply Jensen's inequality to show that  $KL(p, q) \geq 0$ . This is known as the *information inequality*.

c) Use the result of part b to give another proof that the expected length of any prefix code is greater than  $H(X)$ , i.e.:

$$\sum_x p(x)l(x) \geq H(p)$$

Hint: Use the Kraft inequality and set  $q(x) = 2^{-l(x)}$ . Make sure you add a “dummy” element to so that  $q(x)$  is a valid distribution.

Problem 3. A hierarchical hidden Markov model (HMM) has a hierarchy of hidden states. A three layer model has a top level state that changes slowly, a second level state that changes more rapidly and with state transitions depending on the higher level state, and a third level state that changes more rapidly than the second level states with transition probabilities that depend on both higher level states. The state transition probabilities at higher levels do not depend on the lower level states. Furthermore, when a higher level state changes between time  $t$  and  $t+1$  it causes a “reinitialization” of all lower states so that the lower level states at time  $t+1$  do not depend on their values at time  $t$ .

a) Implement a three level hierarchical hidden Markov model as traditional (one-level) hidden Markov model. More specifically, define the state space, emission probabilities and transition probabilities of the one-level hidden Markov model implementing the three level model.

b) Implement a three level hierarchical HMM as a probabilistic context free grammar.