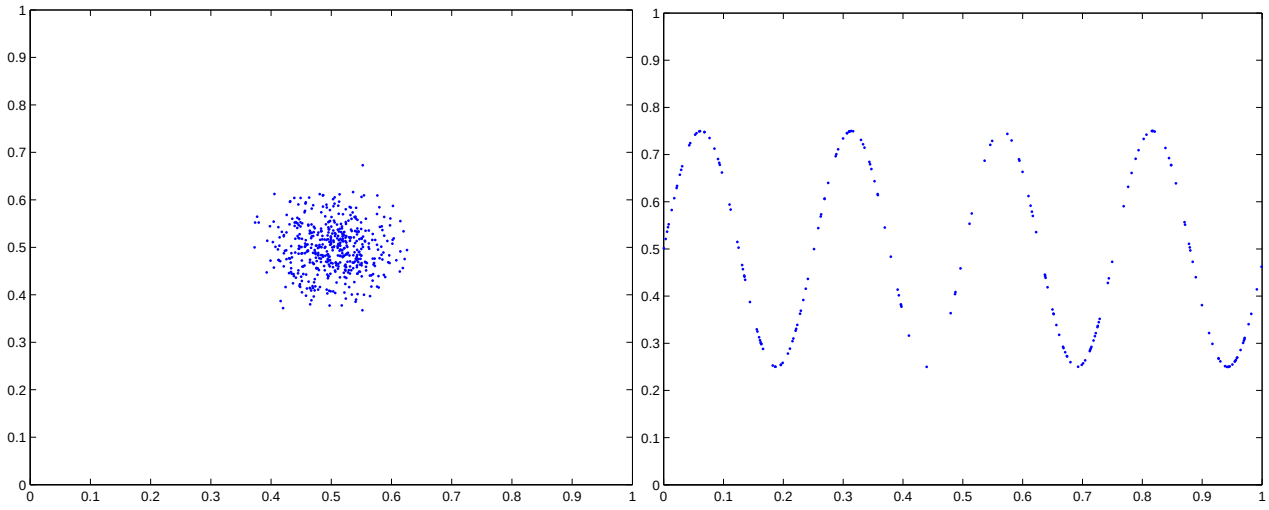
Computer Science 791DD, Learning to See

http://www.cs.umass.edu/∼elm/learning2see/

Exam 2

1. (5 points) The scatter plot on the left shows a sample of 500 points drawn from distribution A. The plot on the right shows a sample of 200 points drawn from distribution B. Assume the support of both distributions is contained within (but not equal to) the unit square. That is, neither distribution has any non-zero density outside of box defined by squares shown. Although one cannot be 100% sure (due to uncertainty of the sample), assess which distribution is likely to have higher differential entropy. Say why.



Answer: First note that because there are more points from distribution A (500) than from distribution B (200), the fact that the neighbor distance appears closer for A does not necessarily mean that distribution A has lower entropy. In addition, one has to be careful when applying the "average log distance technique" with distributions that are not smooth. In particular, the spatial derivatives of distribution B are not smooth since this is effectively a "ridge" distribution with infinitely high density. Why is the density infinitely high for distribution B? Because otherwise the area integral over any part of the density would give zero, since it is infinitely thin.

Several people suggested that because the support of distribution A appeared to be contrained to a smaller rectangular or circular region than that of B, that A might have lower entropy. But this analysis is flawed. The true support of distribution B is actually much smaller than that of distribution A, since distribution B lies on a line. The fact that the line is not straight is not relevant. A few people pointed out correctly that distribution B appears to lie on a one-dimensional manifold (a line) and therefore must have very low entropy (possibly negative infinity).

Thus, the correct answer is that distribution B has the lower differential entropy. In some cases I gave partial credit for the wrong answer or less than full credit for the correct answer, depending upon your reasoning.

2. (25 points) Draw an optimal tree-structured graphical model for the random variables A, B, C, and D given the following information. The meaning of "optimal" should be such that the KL-divergence between the true distribution and the distribution represented by the graph is minimal. Write down all computations that you used to determine the tree-structured graph. You will get partial credit for your correct computations even if you get the tree wrong. If you find an inconsistency in the information below (I don't think that there are any) you will get full credit for the problem. Be careful to note the difference between "H" and "I" in the given information.

   (a) $H(A)$ is 2 bits.
   (b) $H(D|C)$ is 1 bit.
   (c) A is independent of C.
   (d) B can be determined exactly from C.
   (e) D is a uniform distribution over 2^{17} states.
   (f) $I(A,B)$ is 1 **nat**.
   (g) $H(A,D) = 14$ bits.
   (h) $H(B)$ is 8 bits.
   (i) $I(D,B)$ is 7 bits.

   Answer: From the information given, one can conclude the following:

   (a) $I(A,B) < 2$ bits and greater than 0 bits.
   (b) $I(A,C) = 0$ bits.
   (c) $I(A,D) = 5$ bits.
   (d) $I(B,C) = 8$ bits. ($H(B|C) = 0$)
   (e) $I(B,D) = 7$ bits.
   (f) $I(C,D) = 16$ bits.

   From this information, we add edges to the tree in the following order: (C-D), (B-C), and then (A-D).

3. (10 points) You draw a sample from a continuous random variable A of the following numbers:
(14, 15, 10, 8, 0, 9, 64, 32, 16, 12).
You draw a sample from random variable B of the following numbers:
(133, 132, 131, 128, 0, 129, 130, 134).
Using the 1-spacing estimate of entropy, estimate which has higher entropy. You do not need to calculate the exact entropy itself. Just show why you think one entropy is higher than the other.
Answer: (NOTE: I graded this problem somewhat leniently, as I realized it would have made a little more sense if I had given the exact same number of data points in each sample. With the same number of data points, one can merely compare the sum of the log 1-spacings to see which has higher entropy. Since the samples have different numbers of points (10 for A and 8 for 8), one should technically compute the average log 1-spacings and add the log of the number of points. This was not the intent of the problem, however, and by ignoring the fact that the sample sizes were different and just looking at the average log 1-spacings, you still get the correct answer. I gave full credit if you got the idea that you should compare the sum of log 1-spacings or the average of the log 1-spacings, and if you got the comparison in the right direction.)

   1-spacings estimates of entropy are a function of the average of the logs of the 1-spacings. The 1-spacings for variable A are $\{8,1,1,2,2,1,1,16,32\}$. The sum of the logs of these numbers is $3+0+0+1+1+0+0+4+5$=14. The average log 1-spacings is 14/9.

   For variable B, the 1-spacings are $\{128,1,1,1,1,1,1\}$. The sum of the logs of these numbers is $7+0+0+0+0+0+0=7$. The average log 1-spacings is 7/7=1.0.

   Comparing the average log 1-spacings, we see that distribution A has higher differential entropy.

4. (10 points) Explain the idea behind using Independent Components Analysis to do density estimation.

   Answer: In estimating multi-dimensional densities, in the general case, one must have a large number of samples to get accurate density estimates. But in the special case when the components of a distribution are independent, then each marginal can be estimated independently. Since these marginals are lower-dimensional, the density estimates based upon independent marginal estimates are likely to be better. In using ICA for density estimation, the goal is to find a transformation of the data such that the marginals of a distribution can be estimated independently and hence more accurately.

5. (3 points) We read a paper early in the course about the early processing of signals in the visual system of the fly. There were two separate steps in this process. If you think about these steps as a way of modeling the information that is coming into the eye, which of the four semi-parametric models that we discussed in class is best-approximated by the two-step process in that paper?

   Answer: We discussed four different semi-parametric models: product distributions, ICA, tree-distributions, and TCA.

   The first step in the paper was to "decorrelate" or make as independent as possible different measurements in different eye cells. The second process was to model each of these new "independent" signals effectively. This is exactly the approach to density estimation used in Independent Components Analysis, one of the semi-parametric models, as discussed in Problem 4.

6. A two-dimensional distribution is made using the following process. A two-dimensional Gaussian distribution with mean at $(0,0)$ and covariance matrix equal to the identity matrix is multiplied point-wise by an infinite checkerboard that is also centered at the origin. The checkerboard is equal to zero in the black squares and equal to 2 in the white squares.

   (a) (3 points) Is this function C a probability density? Why?

   Answer: Yes, it is a density. Consider the symmetry of the problem. For every part of the Gaussian that gets multiplied by zero (a black square), there is an identical part that gets multiplied by a two (a white square). The black squares remove half of the mass of the original Gaussian and the multiplication by 2 (in the white squares) doubles the mass and returns it to unity.

   (b) (7 points) What is the best product distribution approximation of C and why? Be as specific as you can about the form of the distribution.

   Answer: (NOTE: I should have stated, for this problem, that the "best approximation" was in the sense of KL-divergence. However, I believe everybody correctly assumed that this was the sense in which I meant approximation. Apologies....).

   The best approximation should be the product of the marginals of this "checkerboard Gaussian distribution". But it turns out that the marginals of this checkerboard distribution are exactly the same as the marginals of the original distribution. To see this, think about folding the checkerboard distribution around the x-axis, so that the white checks from the bottom half cover the black checks from the upper half. This "folded" distribution will look exactly like half of the original distribution except that it will have twice the height. It should be obvious that this "half distribution" has, as an x-marginal, a Gaussian distribution, which means the "checkerboard Gaussian" also has a Gaussian for an x-marginal. An analogous argument applies for the y-marginal.

   Thus, the "checkerboard Gaussian" has the same marginals as the original Gaussian distribution. Furthermore, the product of these marginal Gaussians is exactly the original Gaussian distribution, since the components of the original diagonal covariance Gaussian are independent!

   Thus, a zero-mean Gaussian with identity covariance is the best approximation (in the KL-divergence sense) to distribution C.

   (c) (10 points) In general, what is the best product distribution approximation to a joint distribution? Prove this for the two-dimensional case.

   Answer: See next page.

$$\arg\min_{q(x)r(y)} \int\int p(x,y)\log\frac{p(x,y)}{q(x)r(y)}\,dy\,dx \tag{1}$$

$$= \arg\min_{q(x)r(y)} \int\int p(x,y)\log\frac{p(x,y)p(x)p(y)}{p(x)p(y)q(x)r(y)}\,dy\,dx \tag{2}$$

$$= \arg\min_{q(x)r(y)} \int\int p(x,y)\log\frac{p(x,y)}{p(x)p(y)}\,dy\,dx + \int\int p(x,y)\log\frac{p(x)p(y)}{q(x)r(y)}\,dy\,dx \tag{3}$$

$$= \arg\min_{q(x)r(y)} \int\int p(x,y)\log\frac{p(x)p(y)}{q(x)r(y)}\,dy\,dx \tag{4}$$

$$= \arg\min_{q(x)r(y)} \int\int p(x,y)\log\frac{p(x)}{q(x)}\,dy\,dx + \int\int p(x,y)\log\frac{p(y)}{r(y)}\,dy\,dx \tag{5}$$

$$= \arg\min_{q(x)r(y)} \int p(x)\log\frac{p(x)}{q(x)}\,dx + \int p(y)\log\frac{p(y)}{r(y)}\,dy. \tag{6}$$

This last expression is minimized only when $q(x) = p(x)$ and $r(y) = p(y)$, since KL-divergence is minimized (and is equal to 0) only when the distributions match.

It is important to note that we cannot apply this idea directly in (4) because we do not have an expression of the form $\int a\log\frac{a}{b}$. In particular, the $p(x,y)$ does not match the $p(x)p(y)$ in the numerator of the fraction.

7. (10 points) List as many different parameters as you can that people can control (either voluntarily or involuntarily) that affect how light comes into the eye and lands on the retinas. Five or more will get you full credit.

   Answer: Closing eyelids, squinting, tilting head to cause eyebrows or eyelashes to block light, iris changes, blocking light with hands or fingers or hair, focussing or defocussing the eye, rotation of eye, translation or rotation of head, vergence of eyes, blinking to clean eyeball, raising eyebrows to let in more light, etc.

8. (3 points) Describe how hearing (audition) might facilitate learning or signal detection in a particular vision task. You can make up any task you like.

   Answer: Many answers are possible here. I am looking for you to tie together the auditory and visual, rather than discussing them separately.