Applied Information Theory

Midterm Review Sheet

To prepare for the test, you should definitely have completely read and understood the readings from Chapter 2 and Chapter 8 as described on the course web page (in the First Edition of Cover and Thomas, it would be Chapters 2 and 9). I will not expect you to be able to reproduce all parts of all of those chapters, but I expect you to have read them.

On the review sheet and the exam, like in the assignment, I will use the following notation, unless specified otherwise:

- Random variables in capitals: X, Y, Z.
- Sets representing outcomes of a random variable in calligraphic upper case: $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$.
- Sample value of a random variable in lower case x, y.
- Multiple samples of a random variable: $x_1, x_2, ..., x_n$.
- Generic discrete probability distribution $p(x)$.
- Generic continuous probability density $f(x)$.
- Cumulative distribution function of a density $f(x)$: $F(x)$.
- Discrete entropy: $H(X)$ or $H(p(x))$, or for a Bernoulli random variable $H(p)$.
- Differential entropy: $h(X)$ or $h(f(x))$.
- KL-divergence between probability distributions: $D(p(x)||q(x))$ or $D(p||q)$.
- Mutual information between random variables X and Y: $I(X;Y)$.
- Mutual information between random variable X and the joint distribution of Y and Z: $I(X;Y,Z)$.

You should know the following material:

1. The *exact* mathematical definition of discrete entropy.
2. The *exact* mathematical definition of KL-divergence for two discrete probability distributions.
3. The *exact* mathematical definition of mutual information for two discrete random variables
4. The exact definition of differential entropy.
5. KL-divergence and mutual information for continuous random variables.
6. Know the exact statement of Jensen's inequality, and how to apply it.
7. You should be able to apply Jensen's to concave or convex functions, and to know which is which. (Concave function looks like a cave!)
8. I want you to memorize the proof that shows that KL divergence is always greater than or equal to 0. This may sound very "high school" but it will force you to learn several key ideas. The proof is on page 28, Cover and Thomas.

9. IMPORTANT. Know the definition of conditional entropy. In particular, know the difference between $H(X|Y=y)$ and $H(X|Y)$. The first one is an expectation only over X, whereas the second one is an expectation over both X and Y. Understand (and be able to give an example) of why $H(X|Y=y)$ can be *higher* than $H(X)$. Be able to prove that $H(X|Y)$ is *always less than or equal to* $H(X)$. As Cover and Thomas say "conditioning reduces entropy".

10. Be able to show the following directly from the definitions, with at least two intermediate steps:

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(X,Y) && (1)\\ H(X,Y) &= H(X) + H(Y|X) && (2)\\ I(X;Y) &= H(X) - H(X|Y) && (3)\\ & && (4) \end{aligned}$$

11. Know the definition of statistical independence for two or more random variables. For two variables, this is $p(x,y) = p(x)p(y)$.

12. Understand implications of independence of X and Y:

$$\begin{aligned} H(X,Y) &= H(X) + H(Y) && (5)\\ I(X;Y) &= 0 && (6)\\ H(X|Y) &= H(X) && (7)\\ & && (8) \end{aligned}$$

13. Know that a factorial distribution is one with independent components. For example $p(x,y,z) = p(x)p(y)p(z)$ is a factorial distribution.

14. Know the formal definition of a marginal distribution. For $p(x,y)$, the marginal distributions are $p(x) = \sum_x p(x,y)$ and $p(y) = \sum_y p(x,y)$.

15. Know how to prove that of all factorial distributions over X and Y of the form $q(x)r(y)$, the one that minimizes the KL-divergence to $p(x,y)$ is $p(x)p(y)$, i.e., the product of the marginal distributions. This is the *Pythagorean Theorem* for KL divergences. (I did this in class.)

16. Explain alignment of medical images by maximization of mutual information. Explain why alignment by minimizing image differences or maximizing correlation may not work for certain pairs of images.

17. Give 2 ways to estimate the differential entropy of a one-dimensional distribution from a sample. (Parzen windows followed by Monte Carlo and Vasicek (m-spacings) estimator.)

18. Understand the joint alignment algorithm "congealing" which jointly aligns images by minimizing the entropies through the pixels. Why use entropy as a criterion instead of say, the variance? (Answer: the "true" distributions of pixel values, when aligned, might be bimodal, which would have low entropy, but not low variance. Example: MRI images.)

19. Understand the difference between congealing and alignment by maximization of mutual information (AMMI). Answer: in AMMI, only two random variables are defined, one for image A and one for image B. A joint distribution is defined which gives the probability of seeing a *pair* of brightness values in the same location in the two images. It is by maximizing the dependence among the 2 random variables that one aligns the images. In congealing, on the other hand, there is a random variable defined at each pixel location. Each image contains a different sample of the same random variable for a particular location. For example, if each image is 100x100, then there are 10,000 random variables defined. The algorithm aligns the images to minimize the sum of the entropies of *all* the random variables.

20. Understand the probability integral transform. That is, if $f(x)$ is a *probability density function* for a random variable X, and $F(x)$ is its cumulative distribution function, then the **random variable** defined by $F(X)$ is *uniformly distributed* between 0 and 1. You can think of this as the *percentile* random variable. In other words, if I pick a random person from a population, what percentile value do they have, for some random quantity like "age". The answer is, they are just as likely to be at the 13% percentile as in the 47% or 99% percentile. In other words, the percentile of their age is uniformly distributed.

21. Under independent components analysis. Be able to explain what the scatter plot of a factorized distribution should "look like" (answer, a rectangular axis-aligned pattern). Why does ICA work? What does it minimize? What is it minimizing over?

22. Know the Chow-Liu algorithm (This is building a tree graph over random variables by constructing a maximum spanning tree, using the mutual information between nodes as the weight between edges.)

23. Know the independence bound on entropy, and know how to prove it for two variables.

24. What is the maximum differential entropy distribution over an interval? (answer: uniform distribution over that interval.

25. What is the maximum entropy distribution with a fixed variance σ^2? Answer: a Gaussian.

26. Understand why the differential entropy depends upon the units of the underlying probability density function.

27. Differential entropy doesn't change under translation (in one dimension) or under rotation in other dimensions.

28. Understand why we can ignore the joint entropy calculation in mutual information estimation when we are searching over rotations.