Homework assignment #7

Due in class Mon 10/29

Getting comfortable with mutual information (20 points)

1. The harsh weather in North Dakota

A San Diego native, Alice lives in North Dakota and hates winter. She also disapproves of weather apps. Every morning, she looks out of the window and tries to guess how cold it is. After years of observations, Alice compiles a table to help with the guessing. The table shows the joint probability distribution of the ensemble {V,T} = {view from window, temperature}:

		T			
		Miserably cold	Very cold	Cold	Chilly
V	Sunny	1/16	1/16	1/16	1/16
	Cloudy and dry	1/16	1/8	1/32	1/32
	Cloudy and rain	1/8	1/16	1/32	1/32
	Cloudy and snow	1/4	0	0	0

a. What is the joint entropy $H(V,T)$? What are the marginal entropies $H(V)$ and $H(T)$?

b. For each value of v, calculate the conditional entropy $H(T \mid V = v)$.
By definition, this is the entropy of the conditional distribution $P(T \mid V = v)$.
Observe that sometimes, learning V can in fact *increase* the uncertainty in T!

c. Calculate the conditional entropy $H(T \mid V)$ (defined as the average of the above over v).

d. Calculate the conditional entropy $H(V \mid T)$.

e. Calculate the mutual information between V and T.

2. Entropy for continuous variables

The entropy of a discrete probability distribution is $H(\{p_i\}) = -\sum_i p_i \log p_i = -\langle \log p_i \rangle$, where the angular brackets denote averaging with respect to the probability distribution p. This way of writing it naturally generalizes to continuous variables. For a continuous distribution defined by the probability density function $P(x)$, we set: $H(P) = -\int_{-\infty}^{\infty} P(x) \log P(x)\, dx$.

a. Compute the entropy of a Gaussian $P(x) = \frac{1}{\sqrt{2\,\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$. Show that unlike "normal" entropy, the continuous entropy (also called "differential entropy") can be negative.

Thankfully, quantities for which we have intuition, such as mutual information, are defined through a *difference* of entropies, and always turn out positive. Let X be a Gaussian random variable with a large width (i.e. standard deviation) σ_X. Let Y be a "noisy measurement" of X: $Y = X + \xi$, where ξ is a Gaussian random variable with a small width σ_ξ.

b. ξ and X are independent. Without writing any integrals, what is the variance of Y?

c. Show that the probability density function $P_Y(y)$ is given by:

$$P_Y(y) = \int_{-\infty}^{\infty} P_X(x - z)\, P_\xi(z)\, dz$$

This is called the *convolution* of P_X and P_ξ.

d. Show that Y is a Gaussian random variable with the standard deviation you expected ("sum of independent Gaussians is a Gaussian"). **Even if this fails, move to (e).**

e. Compute the mutual information between X and Y. See if it behaves as expected in the limits of small and large σ_ξ/σ_X. (Use the result of (d) even if you skipped the proof.)

3. Converting from bits to dollars **[adapted from MacKay 36.8]**

A horse race involving N horses occurs repeatedly, and you are obligated to bet all your money each time (but you can split your bet between horses in any way you like). Your bet at race n can be represented by a normalized probability vector $\{b_i\}$ multiplied by your money $m(n)$. The bookmaker offers equal odds for each horse; if horse i wins, the respective bet is multiplied by N, and your money becomes $m(n+1) \ = \ N\, b_i\, m(n)$.

a. Calculate the long-term growth of your capital, i.e. the expected value of $\log m(n)$.

b. Show that the *optimal* betting strategy b_i maximizing this growth rate is $b_i = p_i$, and gives an expected growth rate $\lambda_{\max} \ = \ \log N - H(p)$, where H is the entropy function. The bookies' equal odds are appropriate for horses that are equally likely to win. If the wins are *not* equiprobable, and you know the real probabilities – you can make money!

c. You do not know the true probabilities, but estimate them to be q_i and bet accordingly (i.e. with $b_i = q_i$). Show that your capital falls short of the optimal growth by a factor:

$$m(n) = \ m_{\max}(n)\, \mathrm{e}^{-n\, D_{KL}(p\|q)}, \text{ where } D_{KL}(p \parallel q) \equiv -\sum_{\mathrm{i}} p_i \log\frac{q_i}{p_i}.$$

Thus defined, $D_{KL}(p \parallel q)$ is called *the Kullback-Leibler divergence* between probability distributions p and q. It is always positive, vanishing only when $p = q$. It is, however, not symmetric, which is why it is called a "divergence", and not a "distance": $D_{KL}(p \parallel q) \neq D_{KL}(q \parallel p)$.

d. The bookmaker catches on to you and adjusts the winning odds to reflect actual winning probabilities: now if horse i wins, your bet is multiplied by $1/p_i$: $m(n+1) = \frac{b_i\, m(n)}{p_i}$. Show that now the optimal strategy breaks even (the long-term capital growth is 0).

e. You befriend Jack, who works in the stables. Before each race, Jack tells you how happy each of the horses looked that morning. Now you have an edge: the bookies' odds reflect the overall probability $P(i \text{ wins})$, but the probabilities $P\ (i \text{ wins} \mid H_n)$ conditioned on the latest happiness report H_n provide a more accurate estimate of outcomes, allowing you to adjust your bets for each run. Proceeding similarly to (c), show that your winnings grow as follows (careful with the signs!):

$$\log m(n) = \log m_0 + \left\langle D_{KL}\big(P(i \text{ wins}|H) \ \ \| \ \ P(i \text{ wins})\big)\right\rangle_H$$

where $\langle \dots \rangle_H$ denote the averaging over races, and thus over happiness reports.

f. Show that $\left\langle D_{KL}\big(P(X|Y) \ \ \| \ \ P(X)\big)\right\rangle_Y$ is exactly the mutual information $I(X,Y)$.

In other words, the value of extra information is the same in dollars as is it in bits!