CSE 150. Assignment 1 *Summer 2016*

Out: *Tue Jun 28*
Due: *Fri Jul 01 (to box outside CSE 3214 by 10 am)*
Supplementary reading: RN, Ch 13; KN, Ch 1.

1.1 Probabilistic reasoning: Bayes rule

Suppose that 20% of citizens cheat on their taxes. These tax-evading citizens deceive the government by filing forms with incorrect information; in other words, their tax forms are *never* correct.

It turns out, however, that tax forms are very complicated; as a result, sometimes even law-abiding citizens (the other 80% of the population) do not report their income correctly. In particular, studies show that 25% of the forms filed by law-abiding citizens contain honest mistakes.

An audit reveals that a tax form did not correctly report a taxpayers income. What is the probability that this taxpayer was purposefully cheating the government?

1.2 Conditioning on background evidence

It is often useful to consider the impact of specific events in the context of general background evidence, rather than in the absence of information. Denoting such evidence by E, prove the conditionalized version of Bayes rule:

$$P(X|Y,E) \;=\; \frac{P(Y|X,E)P(X|E)}{P(Y|E)}.$$

1.3 Conditional independence

Show that the following three statements about random variables X, Y, and E are equivalent:

$$\begin{aligned} P(X,Y|E) &= P(X|E)P(Y|E) \\ P(X|Y,E) &= P(X|E) \\ P(Y|X,E) &= P(Y|E) \end{aligned}$$

In other words, show that each one of these statements implies the other two. You should become fluent with all these ways of expressing that X is conditionally independent of Y given E.

1.4 Kullback-Leibler distance

Often it is useful to measure the difference between two probability distributions over the same random variable. For example, as shorthand let

$$\begin{aligned} p_i &= P(X=x_i|E), \\ q_i &= P(X=x_i|E') \end{aligned}$$

denote the conditional distributions over the random variable X for different pieces of evidence $E \neq E'$. The Kullback-Leibler (KL) distance between these distributions is defined as:

$$\mathrm{KL}(p,q) = \sum_i p_i \log(p_i/q_i).$$

(a) By sketching graphs of $\log x$ and $x-1$, verify the inequality

$$\log x \leq x - 1,$$

with equality if and only if $x=1$. Confirm this result by differentiation of $\log x - (x-1)$. (Note: all logarithms in this problem are *natural* logarithms.)

(b) Use the previous result to prove that $\mathrm{KL}(p,q) \geq 0$, with equality if and only if the two distributions p_i and q_i are equal.

(c) Provide a counterexample to show that the KL distance is not a symmetric function of its arguments:

$$\mathrm{KL}(p,q) \neq \mathrm{KL}(q,p).$$

Despite this asymmetry, it is still common to refer to $\mathrm{KL}(p,q)$ as a measure of distance between probability distributions.

1.5 Mutual information

Closely related to the Kullback-Leibler distance, the mutual information $I(X,Y)$ between two discrete random variables X and Y is defined as

$$I(X,Y) = \sum_{x,y} P(x,y) \log \left[\frac{P(x,y)}{P(x)P(y)} \right],$$

where the sum is over all possible values of the random variables X and Y. As you will see here, the mutual information provides a *quantitative* measure of conditional dependence.

(a) Prove that the mutual information $I(X,Y)$ is nonnegative. (*Hint:* use the result from the previous problem.)

(b) State a sufficient condition for the mutual information $I(X,Y)$ to vanish.