Homework CMU 17017

Homework 1

1. Information Gain, KL-divergence and Entropy

1. When we construct a decision tree, the next attribute to split is the one with maximum mutual information (a.k.a information gain), which is defined in terms of entropy. In this problem, we'll explore its connection to KL-divergence. The KL-divergence from a distribution $p(x)$ to a distribution $q(x)$ can be thought of as a distance measure from p to q:

$$KL(p||q) = -\sum_x p(x) \log_2 \frac{q(x)}{p(x)}$$

if $p(x) = q(x)$, then $KL(p||q) = 0$. Otherwise, $KL(p||q) > 0$.
We can define mutual information as the KL-divergence from the observed joint distribution of X and Y to the products of their marginals.

$$I(X, Y) = KL(p(x, y) || p(x)p(y))$$

a) show that this definition of mutual information is equivalent to the one given in class. That is, show $I(X,Y) = H(X) - H(X|Y)$ and $I(X,Y) = H(Y) - H(Y|X)$ from the definition in terms of KL-divergence. From this definition, we can easily see that mutual information is symmetric: i.e. $I(X,Y) = I(Y,X)$.

a)

$$\begin{aligned}
KL(p(x,y)||p(x)p(y)) &= -\sum_x \sum_y p(x,y) \log_2 \frac{p(x)p(y)}{p(x,y)} \\
&= -\sum_x \sum_y p(x,y) (\log_2 p(x) + \log_2 (y) - \log_2 (x,y)) \\
&= \sum_x \sum_y p(x,y) (\log_2 \frac{(x,y)}{p(x)} - \log_2 (y)) \\
&= \sum_x \sum_y p(x,y) (\log_2 p(y|x) - \log_2 (y)) \\
&= \sum_x \sum_y p(y|x) p(x) \log_2 p(y|x) - \sum_x \sum_y p(y|x) p(x) \log_2 y \\
&= \sum_x p(x) \sum_y p(y|x) \log_2 p(y|x) - \sum_y \left(\sum_x p(y|x) p(x) \right) \log_2 y \\
&= H(Y|X) - \sum_y p(y) \log_2 y \\
&= H(Y|X) - H(Y)
\end{aligned}$$

b) According to this definition, under what condition do we have that $I(X,Y) = 0$?
when X & Y are independents

2. In the class, we define entropy based on a discrete random variable X. Now consider the case that X is a continuous r.v. with pdf $p(x)$. The entropy is defined as

$$H(X) = -\int p(x) \ln p(x)\, dx$$

Assume $X \sim N(\mu, \sigma^2)$

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp - \frac{(x-\mu)^2}{2\sigma^2}$$

derive $H(X)$

let $-\frac{(x-\mu)^2}{2\sigma^2} = y$

$$H(X) = -\int_{-\infty}^{+\infty} p(x) \ln p(x)\, dx$$

$$= \int_{-\infty}^{+\infty} p(x) \left[\ln \frac{1}{\sqrt{2\pi}\sigma} - \frac{(x-\mu)^2}{2\sigma^2} \right] dx$$

$$= \int_{-\infty}^{+\infty} \ln \frac{1}{\sqrt{2\pi}\sigma} \cdot \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) - p(x) \frac{(x-\mu)^2}{2\sigma^2} dx$$

$$= \ln \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} p(x)\, dx - \int_{-\infty}^{+\infty} -\frac{(x-\mu)^2}{2\sigma^2} p(x)\, dx$$

$$= \ln \frac{1}{\sqrt{2\pi}\sigma} * 1 + \frac{1}{\sigma^2} \int_{-\infty}^{+\infty} p(x)(x-\mu)^2 dx$$

$$E(X-\mu)^2 = \mathrm{Var}(X-\mu) + (E(X-\mu))^2$$

$$= \sigma^2 + 0 = \sigma^2$$

$$= \ln \frac{1}{\sqrt{2\pi}\sigma} + 1$$

2. Bayes' Rules and Point Estimator

2. The Poisson Distr. is a useful discrete distribution which can be used to model the number of occurences of something per unit time. For example, in networking, packets arrival density is often modeled with the Poisson Distribution. If X is a Poisson Distr., i.e. $X \sim Pois(\lambda)$ its probability mass function is

$$P(X=x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

It can be shown that $E(X)=\lambda$. Assume now we have n i.i.d data points from possion(λ) : $D = \{X_1, \cdots, X_n\}$

a) show that the sample mean $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the MLE of λ and it is unbiased $E(\hat{\lambda}) = \lambda$

$$E(\hat{\lambda}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n} E\left(\sum X_i\right) = \frac{1}{n} \sum E(X_i) = \frac{1}{n} \cdot n\lambda = \lambda$$

$$P(X_1, \cdots X_n \mid \lambda) = \prod_{i=1}^{n} P(X_i \mid \lambda)$$

$$\log P(X_1, \cdots X_n \mid \lambda) = \sum_{i=1}^{n} \log P(X_i \mid \lambda)$$

$$= \sum_{i=1}^{n} \left((-\lambda) + \log \lambda^{X_i} - \log X_i! \right)$$

$$= -n\lambda + \log\lambda \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \log X_i!$$

$$\frac{\partial}{\partial \lambda} = -n + \frac{1}{\lambda}\sum_{i=1}^{n} X_i = 0$$

$$\Rightarrow \quad \frac{1}{\lambda}\sum_{i=1}^{n} X_i = n$$

$$\Rightarrow \quad \hat{\lambda} = \frac{\sum_{i=1}^{n} X_i}{n}$$

b) Now let's be Bayesian and puts a prior distribution over λ. Assume that λ follow a Gamma distribution with the parameter (α, β), it's probability density Distr.

$$P(\lambda|\alpha\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

$\Gamma(\alpha) = (\alpha-1)!$. Compute the posterior distribution over λ

$$P(\lambda|D) \propto P(D|\lambda) P(\lambda)$$

$$\propto \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

$$\propto \frac{\lambda^{\sum x_i}}{\prod_{i=1}^{n} x_i!} e^{-\lambda} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

$$\propto \lambda^{\sum x_i + \alpha - 1} e^{-(\beta+n)\lambda}$$

$\sim$ Gamma $(\sum x_i + \alpha, \beta + n)$

c) Derive an analytic expression for the MAP of λ under Gamma (α, β) prior

since $P(\lambda|D) \propto \lambda^{\sum x_i + \alpha - 1} e^{-(\beta+n)\lambda}$

$$\ln P(\lambda|D) \propto (\sum x_i + \alpha - 1) \ln\lambda - (\beta+n)\lambda$$

let $\frac{\partial \ln P(\lambda|D)}{\partial \lambda} = 0$

$$\Leftrightarrow (\sum x_i + \alpha - 1) \cdot \frac{1}{\lambda} - (\beta+n) = 0$$

$$\sum x_i + \alpha - 1 - (\beta+n)\lambda = 0$$

$$\lambda^* = \frac{\sum x_i + \alpha - 1}{\beta + n}$$