Lectures 5: Statistical inference

Statistical Methods for Natural Language Processing Fredrik Engström

February 14, 2011

Lectures 5: Statistical inference

Statistical Methods for Natural Language Processing Fredrik Engström

Summary of lecture 4

- $H[X] = E[\log \frac{1}{p_X}].$
- $H(X, Y) = E[\frac{1}{\log p(x,y)}]$
- $H(Y|X) = \sum_{x} p(x)H(Y|X=x)$
- H(X, Y) = H(X) + H(Y|X)

Mutual information

Definition

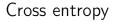
$$I(X;Y) = H(X) - H(X|Y)$$

$$I(X; Y) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p_X(x)p_Y(y)}$$
$$I(X; Y) = E\left[\log \frac{p(x, y)}{p_X(x)p_Y(y)}\right] = E\left[\log \frac{p(x|y)}{p_X(x)}\right]$$

I(X; Y) = 0 iff X and Y are independent

$$I(X;X) = H(X)$$

Lectures 5: Statistical inference



Definition

The cross-entropy between p and q is

$$\sum_{x} p(x) \log \frac{1}{q(x)}.$$

Often p is the true distribution of some variable X and q is a model of p.

Statistical inference

Given a random variable X we say that a sequence of (X_1, \ldots, X_k) of independent random variables, each with the same distribution as X, is a **sample** of X. A sequence of values (x_1, \ldots, x_k) such that $X_i = x_i$ in some experiment is called a **statistical material**. Examples: Dice rolling. Statistical inference: Draw general conclusions (about a population) from a small sample.

Maximum likelihood

- Two bowls of **red** and **white** marbles.
- Bowl 1: 10 red and 10 white.
- Bowl 2: 20 red.

Example: Dice from above. n-grams.

Given some statistical material x_1, \ldots, x_k and some parameters θ .

$$P(x_1,\ldots,x_k|\theta)=\prod P_{\theta}(X_i=x_i).$$

Maximum likelihood estimation (MLE): choose θ to maximize $P(x_1, \ldots, x_k | \theta)$. Smoothing:

- "add one" / Laplace's law: add one to frequency function to get some probabilities even for non appearing tokens.
- "add one half" / Jeffreys-Perks law: add one half to frequency function.

Example: bigrams.

Lectures 5: Statistical inference

Bayesian updating

- Two bowls of **red** and **white** marbles.
- Bowl 1: 10 red and 10 white.
- Bowl 2: 20 red.

Picks bowl 1 with probability $p = \frac{9}{10}$. Example: Hit-and-run.

$$P(H|E) = \frac{P(E|H)}{P(E)}P(H)$$

- **Prior probability**: *P*(*H*)
- Posterior probability: P(H|E)

Choose θ maximizing

$$P(heta|x_1,\ldots,x_k) = rac{P(x_1,\ldots,x_k| heta)P(heta)}{P(x_1,\ldots,x_k)}$$

Bayes decision: Choose s if $P(s|d) \ge P(s'|d)$ for $s' \ne s$.

Lectures 5: Statistical inference

Statistical Methods for Natural Language Processing Fredrik Engström

Summary

- I(X; Y) = H(X) H(X|Y)
- The cross-entropy between p and q is

$$\sum_{x} p(x) \log \frac{1}{q(x)}.$$

- MLE: Maximize $P(x_1, \ldots, x_k | \theta)$
- Needs smoothing with sparse data.
- Bayesian: Maximize P(θ|x₁,...,x_k). Same as maximizing P(x₁,...,x_k|θ)P(θ)