CS 6604: Data Mining Fall 2007

Lecture 18 — October 22, 2007

Lecture: Naren Ramakrishnan *Scribe: Patrick Butler*

1 Overview

In this lecture we first put to rest the solution of how to find the approximation to $n!$ known as Sterling's approximation, using methods gained from statistics. Then we will look at using Bayes's formulation of statistics to create classifiers based on probability that a condition exists given a certain amount of training data.

2 Unfinished Business

Sterling'ss Approximation is given by:

$$n! \approx n^n e^{-n} \sqrt{2\pi n}$$

1. Start with the Poisson distribution

$$P(\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \Bigg| mean = \lambda$$

$x = 1, 2, \ldots, \lambda, \ldots$

We observe that in the region of the mean(λ) Poisson is about normal (mean $= \lambda$ and variance $= \lambda$)

2. Equation-arithmically around the mean the Poisson distribution looks like

$$\begin{aligned} P(\lambda) = \frac{e^{-\lambda}\lambda^\lambda}{\lambda!} &\approx \frac{1}{\sqrt{2\pi\lambda}} * e^{-\frac{\lambda-\lambda}{-2\lambda^2}} \\ \frac{e^{-\lambda}\lambda^\lambda}{\lambda!} &\approx \frac{1}{\sqrt{2\pi\lambda}} * e^0 \\ \lambda! &\approx \sqrt{2\pi\lambda} e^{-\lambda}\lambda^\lambda \Big| \lambda = n \\ n! &\approx n^n e^{-n} \sqrt{2\pi n} \end{aligned}$$

3 Main Section

Last class we talked about posterior and the prior

$P(hyp\|data)$	$\propto$	$P(data\|hyp)$	$\times$	$P(hyp)$
posterior		**likelihood**		**prior**

Assume you are given data and you want to find a hypothesis (e.g. whether the data is random / normal / poisson). Using the Bayesian approach, find the posterior of all of them (maybe the respective values are 0.7, 0.1, and 0.1), and analyze the posterior distribution.

Consider the doctor example with three treatments, three diseases, and three symptoms:

Treatments	Diseases	Symptoms
acupuncture	cancer	S_1
electric shock	attitude	S_7
anti biotics	faking	S_3

Furthermore, assume that all three treatments can be used with all three diseases which will exhibit all three symptoms. If one wanted to know the probability of needing accupuncture given S_1, this could be found using:

$$\begin{aligned} P(accupuncture|S_1) &= \sum_d P(Accupuncture, d|S_1) \\ &= \sum_d P(Accupuncture|d) * P(d|S_1) \end{aligned}$$

But we have made an assumption here!

Recall:

$$\begin{aligned} P(A) &= \sum_B P(A, B) \\ &= \sum_B P(A|B) * P(B) \\ &= \sum_B P(B|A) * P(A) \\ &= P(A) * \sum_B P(B|A) \end{aligned}$$

The above equations hold always, irrespective of the relationship between A and B because then:

$$P(A, B) = P(B) * P(A|B)$$

We can generalize this to three variables in the following way; in general the following holds:

$$P(A, B|C) = P(B|C) * P(A|B, C)$$

With the *additional* information that A & C are conditionally independendent given B, we can conclude:

$$P(A, B|C) = P(B|C) * P(A|B)$$

Going back to the Doctor problem,

$$P(accupuncture|S_1) = \sum_d P(Accupuncture, d|S_1) \quad (1)$$

$$= \sum_d P(d|S_1) * P(accupuncture|d, S_1) \quad (2)$$

$$= \sum_d P(d|S_1) * P(accupuncture|d) \quad (3)$$

For the last step to work, we must assume that $accupuncture$ & S_1 are condionatlly independent given d.

4 Relationship between different modeling paradigms

Recall that the Bayesian approach is to have a distribution as the final answer. The MAP (Maximum A Posteriori) approach picks the hypothesis that yields the maximum in the posterior distribution. The Maximum Likelihood (ML) approach picks the hypothesis that yields the maximum in terms of the likelihood. Note that MAP reduces to ML for a uniform prior.

Consider how a numerical analyst chooses between hypotheses. Given (x,y) numeric data, we can investigate the following two hypotheses:

$$h_1 : y = ax + b$$
$$h_2 : y = ax^2 + bx + c$$

To decide which is best, we can use least squares to find the parameters that yield the least error from each of these hypotheses, i.e., find a, b for h_1 by minimizing:

$$Error = \sum_i (y_i - a * x_i + b))^2$$

or more generally:

$$Error = \sum_i (y_i - f(x_i, a, b, \dots))^2$$

The surprising result is that the least squares hypothesis is actually the same as what we will obtain from a probabilistic interpretation! In particular, the least squares fit is actually the maximum likelihood hypothesis if we assume that the error in the data is normally distributed. For proof, see next class.

Hence, we have the following progression between modeling paradigms: Bayesian → MAP (Maximum A Posteriori) → ML (Maximum likelihood) (when assuming uniform priors) → LS (Least Squares) (when error in data normally distributed).