CS 6604: Data Mining Fall 2007

Lecture 19 — October 24, 2007

Lecture: Naren Ramakrishnan *Scribe: Yong Ju Cho*

1 Overview

In the last lecture we discussed the relationships between different modeling paradigms such as the Bayesian approach, Maximum A Posteriori (MAP) approach, Maximum Likelihood (ML) approach, and the Least-squares (LS) method.

In this lecture we first prove that equivalence of LS and ML under the assumption of normally distributed error. Then, the notions of the naive Bayesian classifier and the Laplace estimate are discussed.

2 Maximum Likelihood and Least Squares

We say that f is a least squares hypothesis if it minimizes $\sum\limits_i [y_i - f(x_i)]^2$.

For instance, when $f(x_i) = ax_i + b_i$, the least square constraints define a quadratic function with bowl shape, thus having a unique (global) minimum. Hence, the line which satisfy the least square constraints is unique.

Let us explore the ML approach with a 'line' hypothesis. Then,

$$\begin{aligned} h_{ML} &= \arg\max_h P(D|h) \\ &= \arg\max_h \Pi_i\, P(d_i|h) \end{aligned}$$

where $P(D|h) = \Pi_i\, P(d_i|h)$ by the assumption of independent events.

If we can assume that the error is normally distributed (around zero), then

$$y_i = f(x_i) + N(0, \sigma^2)$$

Then the ML hypothesis becomes:

$$\begin{aligned} h_{ML} &= \arg\max_h \Pi_i \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(y_i - f(x_i))^2}{2\sigma^2}} \\ &= \arg\max_h \Pi_i\, e^{\frac{-(y_i - f(x_i))^2}{2\sigma^2}} \\ &= \arg\max_h \ln(\Pi_i\, e^{\frac{-(y_i - f(x_i))^2}{2\sigma^2}}) \\ &= \arg\max_h \sum \frac{-(y_i - f(x_i))^2}{2\sigma^2} \\ &= \arg\min_h \sum \frac{(y_i - f(x_i))^2}{2\sigma^2} \end{aligned}$$

As has been shown in the last equation, the least square constraint can be derived from a probabilistic model. As must be clear from the equations, this theory works for not only for a line but also for any hypothesis.

Question: When is LS not appropriate?

Answer: If the error is not normally distributed, the above derivation does not work.

Aside: The data supporting Mendel's laws of inheritance was suspected to be too perfect by Fisher [Fis36]. Because some of Mendel's 'factors' can come from the same chromosome, they may not assort independently. There are still ongoing debates on this issues, e.g., see [Nov04].

3 The Naive Bayes Classifier

The Bayesian approach and MAP approach can have completely different results as in the following example.

Example: Classify a plant as to whether it is poisonous or edible.

Table 1: Hypotheses of edibility

Hypothesis	Predicted class	Posterior probability
h_1	poisonous	0.4
h_2	edible	0.3
h_3	edible	0.3

$$\begin{aligned} P(class|plant) &= \sum_h P(class, h|plant) \\ &= \sum_h P(class|h) \times P(h|plant) \end{aligned}$$

So $P(class = poisonous|plant)$ is:

$$\begin{aligned} P(class = poisonous|plant) &= \sum_h P(class|h) \times P(h|plant) \\ &= 1 \times 0.4 + 0 \times 0.3 + 0 \times 0.3 \\ &= 0.4 \end{aligned}$$

Similarly $P(class = edible|plant)$ is 0.6. The Bayesian approach hence will conclude that the plant is edible. Whereas the MAP hypothesis will conclude that the plant is poisonous.

Question: Which estimation is more credible than the other?

Theoretically, Bayes is right. The Bayes estimate is provably the best estimator.

Let us consider the problem of spam detection. Spam filters classify an email according to features derived from the contents of the message. The features are extracted from the email by the following procedure. First, the content of the email is parsed. Then, stop words such as 'and' and 'is' are removed. After that, stemming of the parsed words are executed. A spam filter finally calculates $P(C|f_1, f_2, f_3, ...f_n)$ to classify a email, where $C = \{S, NS\}$ (spam or non-spam) and f_i are the given features of the email.

Formally, the problem is:

$$\arg \max_{C \in \{S,NS\}} P(C|f_1 \dots f_n)$$

Since $P(C|f_1 \dots f_n) = \frac{P(C|f_1 \dots f_n) \times P(C)}{P(f_1 \dots f_n)}$,

$$\begin{aligned}\arg\max_C P(C|f_1 \dots f_n) &= \arg\max_C P(f_1 \dots f_n|C) \times P(C) \\ &= \arg\max_C \Pi_i P(f_i|C) \times P(C) \text{ (Naive Bayes Assumption)}\end{aligned}$$

(The spam filters might need to retrained regularly because profile of spam changes over time.)

Suprisingly, the Naive Bayes Classifier is one of the best classifiers for detecting spam. In his Ph D thesis [Dom97], Pedro Domingos shows that the naive Bayes classifier performs well under the zero-one loss assumption even if the conditional independence assumption is violated.

Consider the following dataset with three training data examples and two test data examples:

Table 2: Data for the classification

$class$	f_1	f_2
S	marriage	sunny
S	rate	low
NS	grade	low
?	rate	sunny
?	Johny	meet

One of the issues here pertains to how we deal with 'new' words/features such as 'Johnny' in the above table. For newly observed words, $\arg\max_C P(C|f_1 \dots f_n)$ woudl always be zero. Laplace estimation address this issue.

Example: Consider a sample space given by: $forest = \{tiger, elephant, lion, bear\}$

Generalized Laplace estimate of probablity of a event is:

$$\frac{n + p}{N + m},$$

where N is the number of data examples and n is the number of occurrences of the event. Value of p and m depends on how much of the world has been seen so far. For example, Laplace estimate of $P(tiger)$ could

Table 3: Maximum Likelihood vs Laplace

X	$P_{ML}(X)$	$P_{Laplace}(X)$	
$tiger$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{100}$
$elephant$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{100}$
$lion$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{100}$
$bear$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{100}$
$unseen$		$\frac{1}{5}$	$\frac{96}{100}$

be, for instance $\frac{1+0}{4+1}$ or even $\frac{1+0}{4+96}$ as shown in 3. Hence, the Laplace estimate reserves some probability for the unseen instances. The problem with the Laplace estimate is that it is not accurate all the time [GS95], an issue addressed by the Good-Turing estimator. This estimator was invented by Good and Turing during World War II. It is byproduct of researches for breaking the cipher of Enigma. The Good-Turing estimate is shown to be asymptotically optimal[OSZ03].

References

[Dom97] P.M. Domingos. *A unified approach to concept learning*. PhD thesis, University of California at Irvine, Irvine, CA, USA, 1997.

[Fis36] R.A. Fisher. Has Mendel's work been rediscovered? *Annals of Science*, 1(2):115–137, 1936.

[GS95] W.A. Gale and G. Sampson. Good-turing frequency smoothing without tears. *Quantitative Linguistics*, 2(3):217–237, 1995.

[Nov04] C.E. Novitski. Revision of Fisher's analysis of Mendel's garden pea experiments. *Genetics*, 166(3):1139–1140, 2004.

[OSZ03] A. Orlitsky, N.P. Santhanam, and J. Zhang. Always good turing: Asymptotically optimal probability estimation. *Science*, 302(5644):427–431, 2003.