COS 424: Interacting with Data

Lecturer: David Blei & Rob Schapire — Lecture #13
Scribe: Lindsey Poole — March 29, 2007

1 EM Algorithm Review (Dave)

E-Step Compute the posterior distribution of the hidden variables given observations.

M-Step Maximize expected complete log likelihood with respect to model parameters.

1.1 Applications of the EM Algorithm

Language Processing Figure 1 shows a hidden Markov model with noisy observations. The hidden data point determine the next. Instead of a distribution of a cluster, we have a transition matrix between clusters. This is used in language processing. We partition words based on how they are used in articles. EM will cluster words into nouns, verbs, etc.

DNA Data We can also use EM to reconstruct an evolutionary tree and model how DNS evolves over time from the root, as in Figure 2. We could, for instance, reconstruct DNA sequences of dinosaurs.

Face Recognition We can use an EM Algorithm to learn what different parts of the face look like. We divide an image into patches as in Figure 3. We don't know which patch belongs to which face.

Language Translation We could create a English to Spanish translator using an EM Algorithm as in Figure 5

We want to create create a probability distribution, $P(\text{English} \rightarrow \text{Spanish})$, of English to Spanish names.

2 Regression (Rob)

Ok, we're going to change topics now. Previously we've been working with classification algorithms, which categorize objects into particular classes based on their attributes. For instance, we can classify email messages as spam or not spam. However, sometimes you wish to predict a real value quantity based on your observed data. For instance we may wish to predict:

- How much will it rain tomorrow
- How much a stock will go up or down
- How many years a person has to live

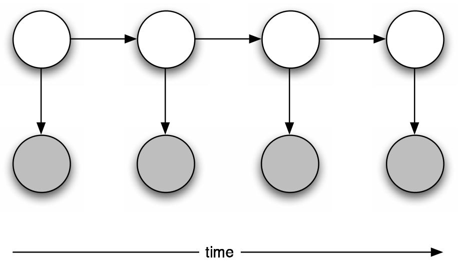

Figure 1: Hidden Markov Model

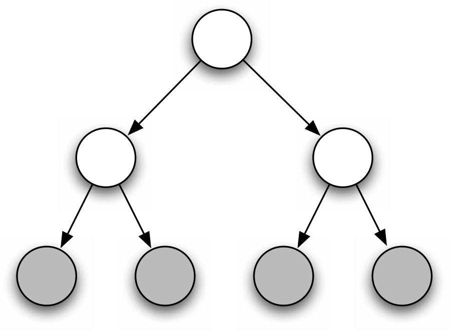

Figure 2: Evolutionary Tree

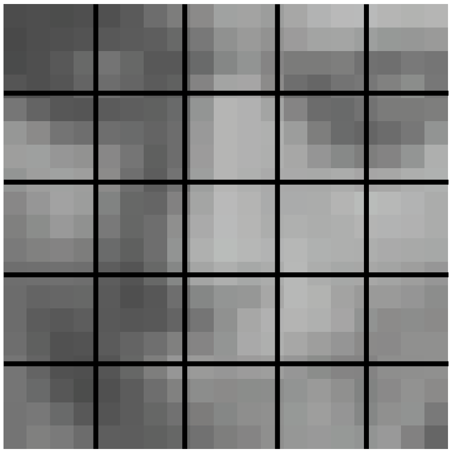

Figure 3: Face Recognition

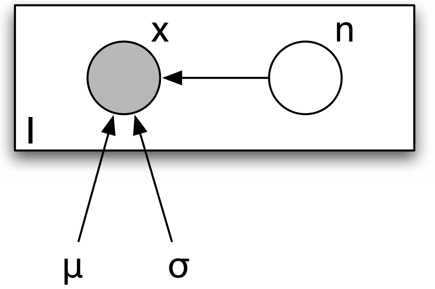

Figure 4: Face Graphical Model

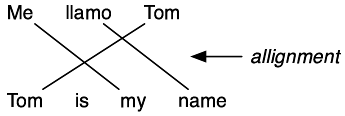

Figure 5: Language Translation

This problem of predicting real values is called regression. Usually our goal is to make predictions that are as accurate as possible. One potential example is to make accurate weather predictions. Occasionally, however, the goal is to uncover some underlying trend or pattern, for instance, the overall trend of the price of a stock.
In order to do this, we are assuming that we have access to a set of $\{x, y\}$ pairs, consisting of an independent and a dependant variable, and we are also assuming that $\{x, y\}$ are independent and identically distributed random variables. The goal is to predict a y value from x, however, we are not trying to predict y exactly; instead, we would like to make a prediction that is as close as possible. So given x, we make a prediction $\hat{f}(x)$ that should be close to y. Alternately, we might want $f(x)$ to be close to

$$f(x) = \underbrace{E[y|x]}_{\text{unknown}}$$

So there are two potential goals we would like to fulfill when estimating $\hat{f}$:

1. we want $|y - \hat{f}(x)|$ to be small
2. we also want $|f(x) - \hat{f}(x)|$ to be small

$f(x)$ is nearly impossible to estimate for most f, as obtaining observations is very difficult. For instance, a person's lifetime can only be measured once, if you are (un)lucky enough to be present to record the events. So for all intents and purposes, $|y - \hat{f}(x)|$ can be measured, $|f(x) - \hat{f}(x)|$ can not, because you can never measure f. Let's change the previous formulae to improve their math niceness:

1. $E_{xy}[(y - \hat{f}(x))^2]$
2. $E_x[(f(x) - \hat{f}(x))^2]$

Theorem 1.

$$\underbrace{E[(y - \hat{f}(x))^2]}_{1} = \underbrace{E[(f(x) - \hat{f}(x))^2]}_{2} + \underbrace{E[(y - f(x))^2]}_{3} \tag{1}$$

If we minimize 1 we also minimize 2. 3 is intrinsic noise.

Proof.

Fix x

$$f = f(x) = E[y|x] = E[y]\hat{f} = \hat{f}(x)$$

We compute both sides:

$$\begin{aligned} LHS &= E[(Y - \hat{f})^2] \\ &= E[y^2 - 2y\hat{f} + \hat{f}^2] \\ &= E[y^2] - 2f\hat{f} + \hat{f}^2. \end{aligned}$$

$$\begin{aligned} RHS &= E[(f-\hat{f})^2] + E[(y-f)^2] \\ &= f^2 - 2f\hat{f} + \hat{f}^2 + E[y^2 - 2fy + f^2] \\ &= f^2 - 2f\hat{f} + \hat{f}^2 + E[y^2] - 2f^2 + f^2. \end{aligned}$$

□

We want to try to minimize 1

Given data $(x_1, y_1), ..., (x_m, y_m)$, we can approximate Equation (1) by

$$\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{f}(x_i))^2. \tag{2}$$

Here, $(y_i - \hat{f}(x_i))^2$ is the loss function, which is also known as the squared error or quadratic loss. We want to minimize $E[loss]$, which is the true loss, risk, or error. To do this, we need to find an $\hat{f}$ that minimizes the empirical loss, as it is an approximation of the true loss.

A different way of thinking about this problem is to assume

$$\begin{aligned} y = f(x) + \quad & \epsilon \quad \text{where} \\ & \epsilon \quad \sim N(0, \sigma^2) \xleftarrow{gaussian} \end{aligned}$$

This is equivalent to

$$Y|X \sim N(f(x), \sigma^2).$$

Then the conditional likelihood of $\hat{f}$, given the x_i's is

$$\prod_{i=1}^{m}\left[\frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(y_i - \hat{f}(x_i))^2}{2\sigma^2}\right)\right]$$

so the log likelihood is

$$-\frac{1}{2\sigma^2}\sum_i \left(y_i = \hat{f}(x_i)\right)^2 - m\ln(\sqrt{2\pi}\sigma)$$

Since σ is a constant, and we're minimizing the summation, we see that minimizing Equation (2) is actually equivalent to maximizing the likelihood under these probabilistic assumptions.

2.1 School District Example

We will now present an example. Our data set for this example consists of school district spending in New Jersey. Figure 6 shows a chart of school district spending, with per district enrollment represented on the x axis, and spending per district on the y axis. Upon closer analysis we discovered that the special education school has much higher funding per student. The question is, how are we to account for this in our spending model? How can we predict how much a school district will spend?

In the following formulae, i is the district number, y_i the spending for district i, and x_i is (in this example anyway) the enrollment. Since y looks really linear, we'll model y by a linear function.

$$\hat{f}(x_i) = wx_i$$

We're trying to minimize

$$\begin{aligned}
\sum_i \left(y_i - \hat{f}(x_i)\right)^2 &= \sum_i (y_i - wx_i)^2 \\
d/dw &= 2\sum_i (x_i(wx_i - y_i)) = 0 \\
w &= \left[\sum_i x_i y_i\right] / \left[\sum_i x_i^2\right].
\end{aligned}$$

So back to our plot, in Figure 7 we see that Newark, the largest school district, seems to be shifting the line up.

What is the obvious way of estimating how much a school district spends on each student? The most obvious way is total dollar amount spent, divided by the total number of pupils:

$$\frac{\sum y_i}{\sum x_i} = \$12098 \text{ per pupil.}$$

This is what is shown by Figure 7.

We now modify our previous log likelihood formula to suit our example. We can assign a different variance, σ_i^2, to each example (x_i, y_i). Then the likelihood is

$$\prod_{i=1}^{m} \left[\frac{1}{\sqrt{2\pi\sigma_i^2}} exp\left(-\frac{(y_i - \hat{f}(x_i))^2}{2\sigma_i^2}\right)\right]$$

and the log likelihood is

$$-\frac{1}{2}\sum_i \left(\frac{y_i - \hat{f}(x_i)}{\sigma_i^2}\right)^2 - \sum_i \ln(\sqrt{2\pi}\sigma_i)$$

We're trying to minimize the first summation.

How should we choose the variances σ_i^2? The total spending for a district y is the sum of the amounts spend on each pupil $y^{(j)}$, so

$$y = y^{(1)} + ... + y^{(x)}.$$

if these random variables are independent, the variance of y is

$$var(y) = var(y^{(1)}) + ... + var(y^{(x)}).$$

Therefore for our example $\sigma_i^2 = cx_i$, so

$$\sum_i \left(\frac{y_i - \hat{f}(x_i)}{cx_i} \right)^2 = \sum_i \left(\frac{y_i - wx_i}{cx_i} \right)^2$$

$$\begin{aligned} d/dw &= \frac{2}{c} \sum_i (wx_i - y_i) = 0 \\ w &= \frac{\sum_i x_i}{\sum_i y_i} \end{aligned}$$

Now we can apply this to our data to obtain a revised spending estimate, as shown in Figure 8.

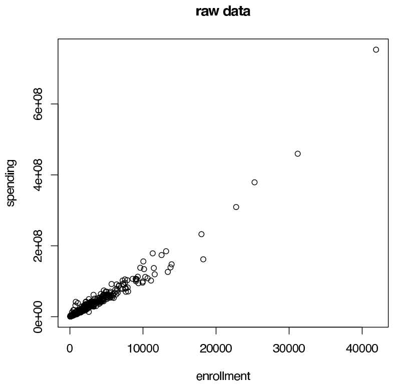

Figure 6: School District Spending

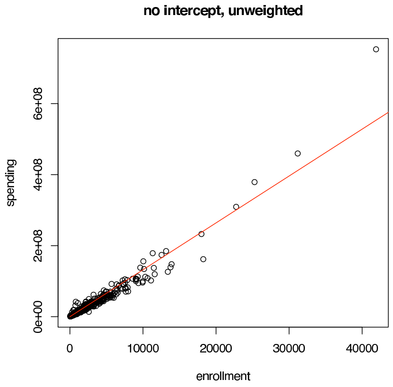

Figure 7: School District Spending

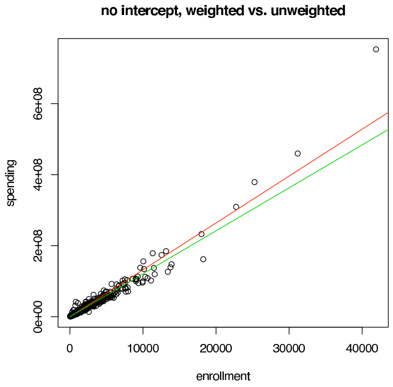

Figure 8: School District Spending