Monday, October 27, 2014 3:14 PM

Hidden Values
Expectation Maximization
Mixture Models

Admin
- Marked Assn 4 due now
- Assn 5 due Wednesday
- Project Proposal due coming Monday
- A6 out tonight, due next Wednesday
- Midterm 2 weeks from today
  (study guide coming soon)

Question From Last time:

Why does bootstrap select ~63% for large N?

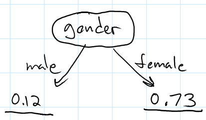
$$p(i \text{ selected at least once}) = 1 - p(\text{not selected } N \text{ times})$$
$$= 1 - (1 - 1/n)^n$$
$$= 1 - 1/e$$
$$\approx 63\%$$

Example of Decision Tree

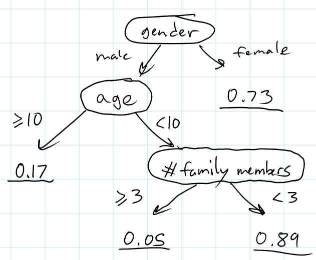
(same as on wikipedia)

$$X = \begin{array}{ccc} \text{Gender} & \text{Age} & \text{\# Family Members} \end{array}$$
$$X = \begin{bmatrix} \text{Male} & 33 & 5 \\ \text{Female} & 10 & 1 \\ \vdots & \vdots & \vdots \end{bmatrix} \qquad y = \begin{bmatrix} \text{`died'} \\ \text{`lived'} \\ \vdots \end{bmatrix}$$

Decision Stump — looks at one variable

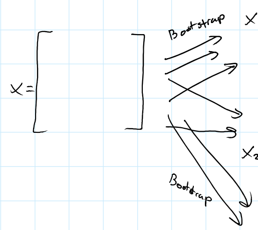


Decision Tree



Random Forests

make a bunch of data sets independently

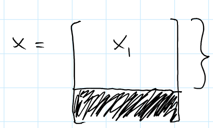


$X_1 = [\ ]$ → fit decision tree to $\{X_1, Y_1\}$

$X_2 = [\ ]$ → fit decision tree to $\{X_2, Y_2\}$

$X_3 = [\ ]$ → fit decision tree to $\{X_3, Y_3\}$

Each decision node is "Random Decision Stump"
- picks random features (subset of features)
- use infogain
- you can prune with "left-out" samples



} ~63% of original data

Hidden Values

- learning when some values are unobserved, missing, hidden, latent

$$
X = \begin{array}{ccc} \text{Gender} & \text{Age} & \text{\# Family Members} \\ \end{array}
$$

$$
X = \begin{bmatrix} \text{Male} & 33 & 5 \\ \text{Female} & 10 & 1 \\ \text{Female} & ? & 2 \\ \text{Male} & 22 & 0 \end{bmatrix} \qquad y = \begin{bmatrix} \text{`died'} \\ \text{`lived'} \\ \text{`died'} \\ ? \end{bmatrix}
$$

Semi Supervised Learning

Idea: Getting labels is expensive, getting unlabeled data is cheap

- can we train on $\{X_L, y_L\}$ and $\{X_U\}$

$$
X_L = \begin{bmatrix} \quad \end{bmatrix} \qquad y_L = \begin{bmatrix} \quad \end{bmatrix}
$$

$$
X_U = \begin{bmatrix} \vdots \end{bmatrix} \qquad y_U = ?
$$

(Very large)

Information inequality:
more data can't hurt!

"Missing at Random" (MAR)

- The fact that it is missing does not depend on the missing value.
- E.g. Digit Classification

$\boxed{2} \to 2.$

$\boxed{3} \to 3.$

- missing random pixels: MAR
- hide the labels of all the "2"s examples (<u>not</u> MAR)
- hide the top half of every digit

- If not MAR, you need to model WHY data is missing.

Approach #1

1. Imputation: replace ?'s with most likely value
   ie. guess the age if its missing
2. Fit model with "imputed values"
3. Can impute missing data with new, fitted model to try and be better

"hard - EM"

Probabilistic Approach

Notation X: observed variables H: hidden variables

$P(X) = \sum_h p(X, H=h)$ (integral if h is a continuous random variable)

Eg. SSL

$p(\bar{Y}_L, X_L, X_U) = \prod_{i=1}^{N} p(y_i, x_i) \prod_{j=1}^{T} \left(\sum_{y_j} p(y_j, x_j) \right)$

Problem: We assume $-\log p(X, H)$ is "nice" (closed-form, convex)

Maximize $p(x)$:

$\log(p(x)) = \log\left(\sum_h p(X, H=h)\right)$

if probability is exponential, we have familiar log-sum-exp

Minimizing: $-\log(p(x)) = \underbrace{-\log\left(\sum_h p(X, H=h)\right)}_{\text{not convex}}$

- problem is the summation inside log

side note:

$\log(1 + \exp(w^T x))$

$\log(\exp(1) + \exp(w^T x))$ "convex"

Expectation Maximization

Local Optimizer when $\log p(X, H)$ is "nice" with parameters θ

Problem: $\max_\theta \; p(X \mid \theta)$

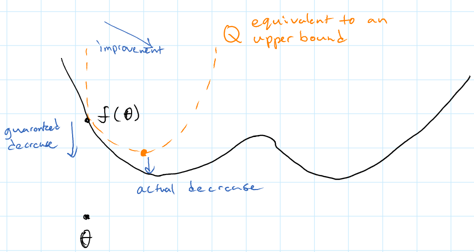
$E_X[f(x)] = \sum_x p(x) f(x)$ $E_{X|Y}[f(x)] = \sum_x p(x|y) f(x)$

Iterations θ^t

"E-Step": Define $Q(\theta \mid \theta^t) = E_{H|X,\theta}\left[\log p(X, H \mid \theta)\right]$

"M-step": $\theta^{t+1} = \underset{\theta}{\operatorname{argmax}} \; Q(\theta \mid \theta^t) = \sum_h \underbrace{p(h \mid X, \theta^t)}_{\alpha_h^t} \underbrace{\log p(X, h \mid \theta)}_{\text{`nice'}}$

(fixed)

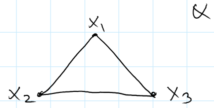


Theorum

$\log(p(x \mid \theta^{t+1})) - \log p(X \mid \theta^t)$

$\geq Q(\theta^{t+1} \mid \theta^t) - Q(\theta^t \mid \theta^t)$

"Convex Combination"



$$\alpha_1 x_1 + \alpha_2 x_2 + \cdots \alpha_n x_3$$

$$\alpha_i \geq 0 \quad \sum_i \alpha_i = 1$$

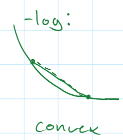
If f is convex,

$$f\left(\sum_i \alpha_i x_i\right) \leq \sum_i \alpha_i f(x_i)$$

$$-\log p(x \mid \theta) = -\log \sum_h p(X, h \mid \theta)$$

multiply & divide by α_h

$$= -\log\left(\sum_h \alpha_h \frac{p(X, h \mid \theta)}{\alpha_h}\right)$$

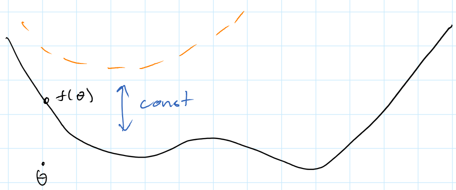


$$\leq -\sum_h \alpha_h \log \frac{p(X, h \mid \theta)}{\alpha_h}$$

$$= -\sum_h \alpha_h \log p(x, h \mid \theta) + \underbrace{\sum_h \alpha_h \log \alpha_h}_{-H(\alpha)}$$

$$= -Q(\theta \mid \theta^t) + \text{const}$$

negative entropy



$$p(x \mid H, \theta) = \frac{p(X, H \mid \theta)}{p(H \mid x, \theta)}$$

$$-\log(p(x \mid \theta)) = -\log(p(X, H \mid \theta)) + \log(p(H \mid X, \theta))$$

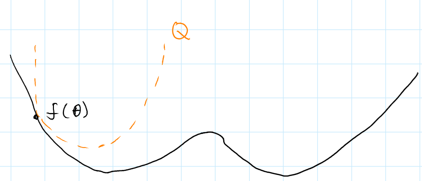
take expectation:

$$\sum_h \alpha_h (-\log p(x \mid \theta^t)) = \sum_h \alpha_h \left(-\log p(X, H \mid \theta^t) + \log p(H \mid X, \theta^t)\right)$$

$$= \sum_h E \, \alpha_h \left[\log p(X, H \mid \theta^t) + \alpha_h \underbrace{\log p(H \mid X, \theta^t)}_{\alpha_h}\right]$$

$$= Q(\theta^t \mid \theta^t) + \text{const}$$

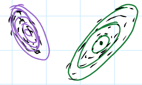
"it does touch the point"



Mixture Models

- Recall fitting Gaussians:

multiple classes: GDA:



- Want to fit probalistic model
  - don't have labels
  - but it would be easy if we did!
  - motivates mixture models

Let $z_i \in \{1, 2, \cdots k\}$

$$p_k(x_i \mid \theta) = \underbrace{p(x_i \mid z_i = k, \theta)}_{\text{"Gaussian"}}$$

z_i is a "latent" variable   never observe, but would be nice to know

$$p(x_i \mid \theta) = \sum_{k=1}^{K} p(x_i, z_i = k \mid \theta)$$

$\sum_i$ for full data set

$$= \sum_{k=1}^{K} \underbrace{p(x_i \mid z_i = k, \theta)}_{\text{~~Gaussian for cluster~~ k}} \underbrace{p(z_i = k \mid \theta)}_{\text{probability that data point belongs to mixture 'k'}}$$

K-means: hard-EM
with Gaussian
with $\Sigma_k = I$

- usually, mixture models are fit with EM

- if one class is moe complicated, can use mixture of a Guassian to get rid of unimodal problem