Hidden values
Expectation Maximization
Mixture Models

- Marked A4: now
- A5: Wednesday
- PP: Monday
- A6: out tonight due next Wed.
  - MT: 2 weeks from today
    (study guide coming soon)

Why does bootstrap select \( \approx 63\% \) for large \( N \)?

\[ p(\text{\textit{i} selected at least once}) = 1 - p(\text{not selected } \text{i} \text{ times}) \]
\[ = 1 - (1 - \frac{1}{n})^n \]
\[ \approx 1 - \frac{1}{e} \]
\[ \approx 63\% \]
Example of decision tree

\[ X = \begin{bmatrix}
\text{Gender} & \text{Age} & \text{#family members} \\
\text{Male} & 33 & 5 \\
\text{Female} & 10 & 1 \\
\vdots & \vdots & \vdots \\
\end{bmatrix} \]

\[ Y = \begin{bmatrix}
\text{died} \\
\text{lived} \\
\vdots \\
\end{bmatrix} \]

Decision Stump

Decision tree

age

- \leq 10
- > 10

\#family members

- \leq 2
- > 2
Random Forests

$X_1 = \ldots$

Fit decision tree to $X_1, \ldots$

$X_2 = \ldots$

Fit decision tree to $X_2, \ldots$

$X_3 = \ldots$

Fit decision tree to $X_3, \ldots$

Each decision node is "random decision stump" - picks random features - use info gain - you can prune with "left-out" samples

$X = [X_i, X_j]$ (approx 63% of original data)
Hidden values
Expectation Maximization
Mixture Models

Hidden Values
- Learning when some values are unobserved, missing, hidden, latent.
Example of decision tree

\[ X = \begin{bmatrix}
  \text{Gender} & \text{Age} & \# \text{family members} \\
  \text{Male} & 33 & 5 \\
  \text{Female} & 10 & 1 \\
  \text{Female} & \text{?} & 2 \\
  \text{Male} & 22 & 0 \\
\end{bmatrix} \quad \text{\textit{y} = \begin{bmatrix}
  \text{died} \\
  \text{lived} \\
  \text{died} \\
  \end{bmatrix}} \]

Semi-Supervised Learning

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

Can we train on \( X_L \) and \( X_U \)?
"Missing at Random" (MAR)

- the fact that it is missing does not depend on the missing value.

- E.g., digit classification
  
  \[
  \begin{align*}
  2 & \rightarrow "2" \\
  3 & \rightarrow "3"
  \end{align*}
  
  missing random pixels: MAR
  
  hide the labels of all the "2" examples. (not MAR)

  hide the top half of every digit: MAR

- If not MAR, you need to model why data is missing.
Approach #1

1. Imputation: replace "?" with most likely value.
2. Fit model with "imputed" values.

"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 rv)

Eg. SSL:

$$p(\tilde{y}_L, x_s, x_u) = \prod_{i=1}^{N} p(y_i, x_i) \prod_{j=1}^{T} \left( \sum_{x_0} p(y_0, x_0) \right)$$
Problem:

Assume $-\log P(X, H)$ is "nice"
(closed-form convex)

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

not convex

"problem is $\sum$ inside log"

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

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

- local optimizer when \( \log p(X, H) \) is "nice" with parameters \( \theta \).

Problem: \( \max \theta \ p(X|\theta) \)

Iterations: \( t \)

"E"-step: Define \( Q(\theta|\theta^t) = \mathbb{E}_{H|X, \theta^t} \left[ \log p(X, H|\theta) \right] \)

"M"-step: \( \theta^{t+1} = \arg\max \theta Q(\theta|\theta^t) = \sum_{h} p(h|X, \theta^t) \log p(X, h|\theta) \) "nice"
Maximization Models

Theorem:

\[ \log p(X|\theta^{t+1}) - \log p(X|\theta^t) \geq Q(\theta^{t+1}|\theta^t) - Q(\theta^t|\theta^t) \]
"Convex Combination"

\[ \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \cdots + \alpha_n x_n \geq 0 \leq \alpha_i = 1 \]

If \( f \) is convex,

\[ f\left( \sum \alpha_i x_i \right) \leq \sum \alpha_i f(x_i) \]
Problem: \( \max_{\theta} p(X | \theta) \)

Iterations: \( \theta^t \)

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

"M"-step: \( \theta^{t+1} = \arg \max_{\theta} Q(\theta | \theta^t) = \sum_{h \in \mathcal{H}} p(h \mid X, \theta^t) \log p(X, h | \theta) \)

"nice"
\[-\log p(X | \Theta) = -\log \sum_h p(X, h | \Theta)\]

\[= -\log \left( \sum_h \alpha_h p(X | h | \Theta) \right)\]

\[\leq -\sum_h \alpha_h \log p(X | h | \Theta)\]

\[= -\sum_h \alpha_h \log p(X | h | \Theta) + \sum_h \alpha_h \log \alpha_h\]

\[= -Q(\Theta | \Theta^t) + \text{const.}\]

\[-H(\alpha)\]
Mixture Models

Recall fitting Gaussians:

Want to fit probabilistic model:
- don't have labels
- but it would be easy if we did
- motivates "mixture" models
Let \( z_i \in \{1, 2, \ldots, k \} \).  

\[ p_k(x_i | \theta) = p(x_i | z_i = k, \theta) \]

"Gaussian"

\( z_i \) is a "latent" variable.

\[ p(x_i | \theta) = \sum_{k=1}^{K} p(x_i | z_i = k | \theta) p(z_i = k | \theta) \]

\( k \)-means: hard-EM  
with Gauss  
with \( \xi_{ik} = 1 \)