Other SSL approaches
Bayesian learning
Conjugate priors

A4: pick at end of class?

A5: due now, marked version due Nov 6
PP: due Monday
A6: due 'friday of, => tutorials this week
MT: November 10

Marking

Cross-validation with regularization:

full dataset:
\[ \sum_{i=1}^{n} f(y_i, w^T x_i) + \frac{1}{2} \| w \|_2^2 \]

Cross-validation fold:
\[ \sum_{i=1}^{n_{cv}} f(y_i, w^T x_i) + \frac{1}{2} \| w \|_2^2 \]
relative of weights are wrong

solutions
just use w from one fold
multiply 1 by \[ q \]
and train on full data.
Corrections to EM analysis

We showed $\log p(X|\theta) \geq Q(\theta|\theta^t) + H(p(H|X,\theta^t))$

Part 2:

\[
p(X|\theta^t) = \frac{p(X, H|\theta^t)}{p(H|X,\theta^t)}
\]

(this was correct)

\[
p(X, H|\theta^t) = p(H|X,\theta^t)
\]

- take log, take expectation...

\[
\sum_h p(h|X,\theta^t) \log p(X|\theta^t) = \sum_h p(h|X,\theta^t) \log p(X, H|\theta^t)
\]

\[
- \sum_h p(h|X,\theta^t) \log p(H|X,\theta^t)
\]

\[
\log p(X|\theta^t) \leq \sum_h p(h|X,\theta^t) = Q(\theta^t|\theta^t) + H(p(H|X,\theta^t))
\]

\[
\Rightarrow \log p(X|\theta) - \log p(X|\theta^t) \geq Q(\theta|\theta^t) - Q(\theta^t|\theta^t) + H(-)
\]
Approaches to SSL

1. EM (for generative models only)

2. Co-training
   - split features into 'views'
     (classifying webpages → text, hyper)
   - train a classifier on each view

3. Entropy regularization
   - penalize entropy of \( p(y|x^u|x_{ SVM}) \)
   - want labels to be non-random.

4. Transductive: "just say no"
   SVM
5. Graph-based SSL

- use features or relationship between $x_i$ to define a graph.
- graph has a weight $w_{ij}$, how much we want $y_i$ and $y_j$ to agree.
Problems with MAP estimation

- Does MAP make the right decision?

\[ H = \{ h_1, h_2, h_3, h_4 \} \]

- Missing notion of risk

\[ p(h_1 | D) = 0.25 \quad p(h_2 | D) = 0.3 \quad p(h_3 | D) = 0.25 \quad p(h_4 | D) = 0.7 \][]
Learning principles

ML: $\hat{h} = \arg\max_h p(O|h)$

predict using $\arg\max_h p(O|h)$

MAP $\hat{h} = \arg\max_h p(h|D) \propto p(O|h)p(h)$

predict using $\hat{h} = \arg\max_h p(O|h)$

Bayesian: work w/ full posterior $p(h|D) = \frac{p(O|h)p(h)}{p(O)} = \frac{p(O|h)p(h)}{\sum p(O|h)p(h)dh}$

predict by integrating over "hidden" parameters

$p(h|D) = \sum \int \hat{h} = \int \sum p(\hat{h}|h)p(h|D)p(h)dh$
Example: Coin flipping.

Bernoulli likelihood

\[ p(X = 'H' | \theta) = \theta \]
\[ p(X = 'T' | \theta) = (1 - \theta) \]
\[ p(X | \theta) = \theta^x (1-\theta)^{1-x} \]

Beta prior on \( \theta \):

Uniform: Beta(5, 1)

\[ \theta \sim \text{Beta}(\alpha, \beta) \]
\[ p(\theta | \alpha, \beta) = \frac{\theta^{\alpha-1} (1-\theta)^{\beta-1}}{B(\alpha, \beta)} \]

\text{'beta' function:} \quad B(\alpha, \beta) = \frac{\Gamma(\alpha) \Gamma(\beta)}{\Gamma(\alpha + \beta)}

\[ \int p(\theta | \alpha, \beta) d\theta = 1 \]
\[ \int \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} d\theta = 1 \]
\[ \int \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} d\theta = B(\alpha, \beta) \]
Posterior

Assume we observe 'HHHH'

\[ p(\theta | HHHH) = \frac{p(HHHH | \theta) p(\theta)}{p(HHHH)} \]

\[ = \frac{p(HHHH | \theta) p(\theta)}{p(HHHH)} \]

\[ = \frac{\theta^{3-a-1} (1-\theta)^{b-1} p(HHHH | \theta) p(\theta)}{p(HHHH)} \]

\[ = \frac{\theta^{(3+a)-1} (1-\theta)^{b-1}}{p(HHHH)} \]

\[ \theta | HHHH \sim Beta(3+a, b) \]
MLE: \( \theta = \frac{N_1}{N} = \frac{3}{3} = 1 \)

MAP: \( \theta = \frac{\alpha - 1}{\alpha + b - 2} = \frac{(3 + a) - 1}{(3 + a + b) - 2} = \frac{2}{3} = 1 \)

mean of posterior:

\[ \frac{\alpha}{\alpha + b} = \frac{(3 + a)}{(3 + a + b)} = \frac{4}{5} \]

\( \approx 80\% \) heads.

\( \alpha = 3, \ b = 3 \): like we have \( \frac{3}{2} \) HTHTHTT before we see data

\( \alpha = 0.1, \ b = 0.1 \)

\( \alpha = 100, \ b = 1 \)
\[ p(H^1 | HHHH) = \sum_{\theta} p(H^1 | \theta) p(\theta | HHHH) \]
\[ = \sum_{\theta} \text{Ber}(H | \theta) \text{Beta}(\theta | 13 + a, b) d\theta \]
\[ = \text{EC} \text{Beta}(\theta | 13 + a, b) \]
\[ = \frac{3 + a}{3 + a + b} \]