Other SSL approaches
Bayesian Learning
Conjugate priors

Admin
- AS: due now
- Project Proposal: due monday
- A6: due wednesday => is a tutorial this week
- Midterm: November 10 (or hand in thurs/friday during tutorials)

Marking
- don't be a troll!
- give half marks!
- give Scott full marks!

Cross-Validation with regularization:

full dataset:

$$\sum_{i=1}^{N} f(y_i, w^T x_i) + \frac{\lambda}{2}\|w\|^2$$

Cross Validation Fold:

$$\sum_{i=1}^{9N/10} f(y_i, w^T x_i) + \frac{\lambda}{2}\|w\|^2$$

10x Smaller

relative of weights are wrong -solutions?

- just use w from one fold
- multiply λ by $\frac{10}{9}$ and train on full data (change scale)

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) $p(x,H|\theta) = p(H|x,\theta)\,p(x|\theta)$

$$\sum_h p(h|x,\theta^t)\log p(x|\theta^t) = \sum_h p(h|x,\theta^t)\log p(x,h|\theta^t) - \sum_h p(h|x,\theta^t)\log p(h|x,\theta^t)$$

$$\log p(x|\theta^t)\boxed{\sum_h p(h|x,\theta^t)} = Q(\theta^t|\theta^t) + H(p(H|x,\theta^t))$$

$\boxed{\sum_h p(h|x,\theta^t) = 1}$

$$\Rightarrow \log(x|\theta) - \log p(x|\theta^t) \geq Q(\theta|\theta^t) - Q(\theta^t|\theta^t) \pm H(\cdot)$$

entropy

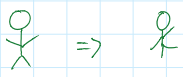
Supervised Latent Variable Models

1. Generative model with mixture for $p(x_i|y_i)$
2. Mixtures of classifiers (mixture of experts)
3. Latent SVM

$$\min_w \frac{\lambda}{2}\|w\|^2 + \sum_{i=1}^{N} \max\{0, 1 - y_i f_w(x_i)\}$$

latent / standard

$$\max_{h \in H}\{w^T \phi(x_i, h)\} \qquad w^T \phi(x_i)$$



deformable part models

End of Scope for Midterm

Approaches to SSL (non-convex world)

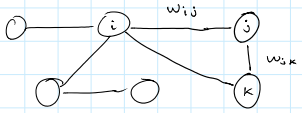
1. EM (for generative models only)

2. Co-training
   - split features into "views"
     classifying webpage: → text on webpage, → hyperlinks
   - train a classifier on each view
   - add high confidence (from either view) unlabeled examples to labeled

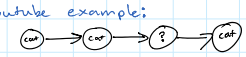
3. Entropy regularization
   - penalize entropy of $p(y_U \mid X_U, w)$
   - want labels to be non-random

4. Transductive SVMs: "Just say no" (it's bad)

5. Graph-based SSL
   - use features or relationship between x_i to define a graph
   - graph has weight w_{ij}, how much we want y_i and y_j to agree



youtube example:



watching labeled videos of cats in a row, the missing labled elemet is also probably a cat video

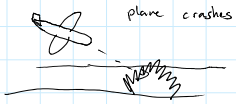
Problems with MAP estimation

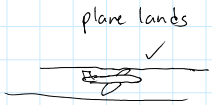
- Does MAP make the right decision?

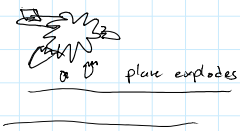
$H = \{h_1, h_2, h_3, h_4\}$

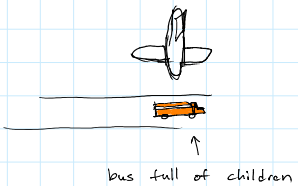
h: hypothesis
D: data
H: hypothesis space

$p(h_1 \mid D) = 0.25$ $p(h_2 \mid D) = 0.3$ $p(h_3 \mid D) = 0.25$ $p(h_4 \mid D) = 0.2$









$p(h_2 \mid D) = 0.3$ (MAP) $p(\neg h_2 \mid D) = 0.7$

$p(\text{not live} \mid D) = p(h_1 \mid D) + p(h_3 \mid D) + p(h_4 \mid D) = 0.7$

if we want to live, MAP solution doesn't exactly represent what we should do

Learning Principles

ML: $\hat{h} = \underset{h}{\text{argmax}}\ p(D \mid h)$ train

point estimate (one $\hat{h}$)

predict using $\underset{\hat{D}}{\text{argmax}}\ p(\hat{D}, \hat{h})$ test

hidden: $\hat{h} = \underset{h}{\text{argmax}} \sum_{z} p(h, z \mid D)$ ← hidden

bagging: $\hat{h} = \sum_{D_i \in \text{bootstrap}(D)} \underset{h}{\text{argmax}}\ p(h \mid D)$

MAP: $\hat{h} = \underset{h}{\text{argmax}}\ p(D \mid h) \propto p(D \mid h)\, p(h)$

predict using $\underset{\hat{D}}{\text{argmax}}\ p(\hat{D}, \hat{h})$

Bayesian: work with full posterior $p(h \mid D) = \frac{p(D \mid h)\, p(h)}{p(D)} = \frac{p(D \mid h)\, p(h)}{\int p(D \mid h) p(h)\, dh}$
(not a "point" estimate)
predict by integrating over "hidden" parameters

$$p(\hat{D} \mid D) = \int_H p(\hat{D}, h \mid D)\, dh$$

$$= \int_H p(\hat{D} \mid h, D)\, p(h \mid D)\, dh$$

$$= \int_H p(\hat{D} \mid h)\, p(h \mid D)\, dh$$

Example: Coin Flipping

Bernoulli Likelyhood:

θ ← probability it comes up heads

$$p(X = \text{'H'} \mid \theta) = \theta$$

$$p(x = \text{'T'} \mid \theta) = (1 - \theta)$$

$$p(X \mid \theta) = \theta^{I(x=\text{'H'})} (1-\theta)^{I(x=\text{'T'})}$$

Beta Prior on θ:

$$\theta \sim \text{Beta}(a, b)$$

$$p(\theta \mid a, b) = \frac{1}{B(a,b)} \theta^{a-1} (1-\theta)^{b-1} \propto \theta^{a-1} (1-\theta)^{b-1}$$

"Beta" function: $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$

uniform: Beta(1, 1)

Posterior:

- assume we observe 'H H H'

$$p(\theta \mid \{HHH\}) \propto \frac{p(\{HHH\} \mid \theta)\, p(\theta)}{p(\{H\,H\,H\})}$$

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

$$= \frac{\theta^3\, \theta^{a-1} (1-\theta)^{b-1}}{p(\{H\,H\,H\})}$$

$$= \frac{\theta^{(3+a)-1} (1-\theta)^{b-1}}{p(\{H\,H\,H\})}$$

Note: we hid $p(\{HHH\} \mid a, b) = B(3+a, b)$ "Marginal likelyhood"

$$\theta \mid \{HHH\} \sim \text{Beta}(3+a, b)$$

MLE: $\theta = \frac{N_1}{N} = \frac{3}{3} = 1$

MAP: $\theta = \frac{\alpha - 1}{\alpha + b - 2} = \frac{3 + a - 1}{3 + a + b - 2} \xrightarrow{\text{under uniform}} \frac{3}{3} = 1$

Mean of posterior: $\frac{\alpha}{\alpha + b} = \frac{3 + a}{3 + a + b} \xrightarrow{\text{uniform}} \frac{4}{5}$ $\approx$ 80% heads

$a = 3,\ b = 3$: like we have $\{HTHT\}$ before we see data

"expectations" on how it will behave

$$p(\hat{H} \mid \{HHH\}) = \int_0^1 p(\hat{H} \mid \theta)\, p(\theta \mid \{HHH\})\, d\theta$$

$$= \int_0^1 \underbrace{\text{Ber}(\hat{H} \mid \theta)}_{\theta}\ \text{Beta}(\theta \mid 3 + a, b)\, d\theta$$

$$= E[\text{Beta}(\theta \mid 3 + a, b)]$$

$$= \frac{3 + a}{3 + a + b}$$ (not generally the same as posterior mean)