Other SSL approaches

Bayesian learning

Conjugate priors

A4: pick at end of class?

A5: due now, marked version due Wed

PP: due Monday

A6: due Friday of next week => tutorials this week

MT: November 10

(*) Marking

Cross-validation with regularization:

full dataset:

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

cross-validation fold:

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

relative of weights are wrong

solutions

just use w from one fold

multiply λ by $\frac{10}{9}$ 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: $$\boxed{p(X|\theta^t) = \frac{p(X,H|\theta^t)}{p(H|X,\theta^t)}}$$ (this was correct)

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

- take log, take expectation....

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

$$\log p(X|\theta^t) \overbrace{\boxed{\sum_h p(h|X,\theta^t)}}^{=1} = Q(\theta^t|\theta^t) + H(p(H|X,\theta^t))$$

$\Rightarrow$

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

App

1.

2.

Approaches to SSL

1. EM (for generative models only)

2. Co-training

   - split features into 'views'

     (classifying webpages → text
     → hyp

   - train a classifier on each view

   - add high confidence unlabeled examples to labeled

3. Entropy regularization

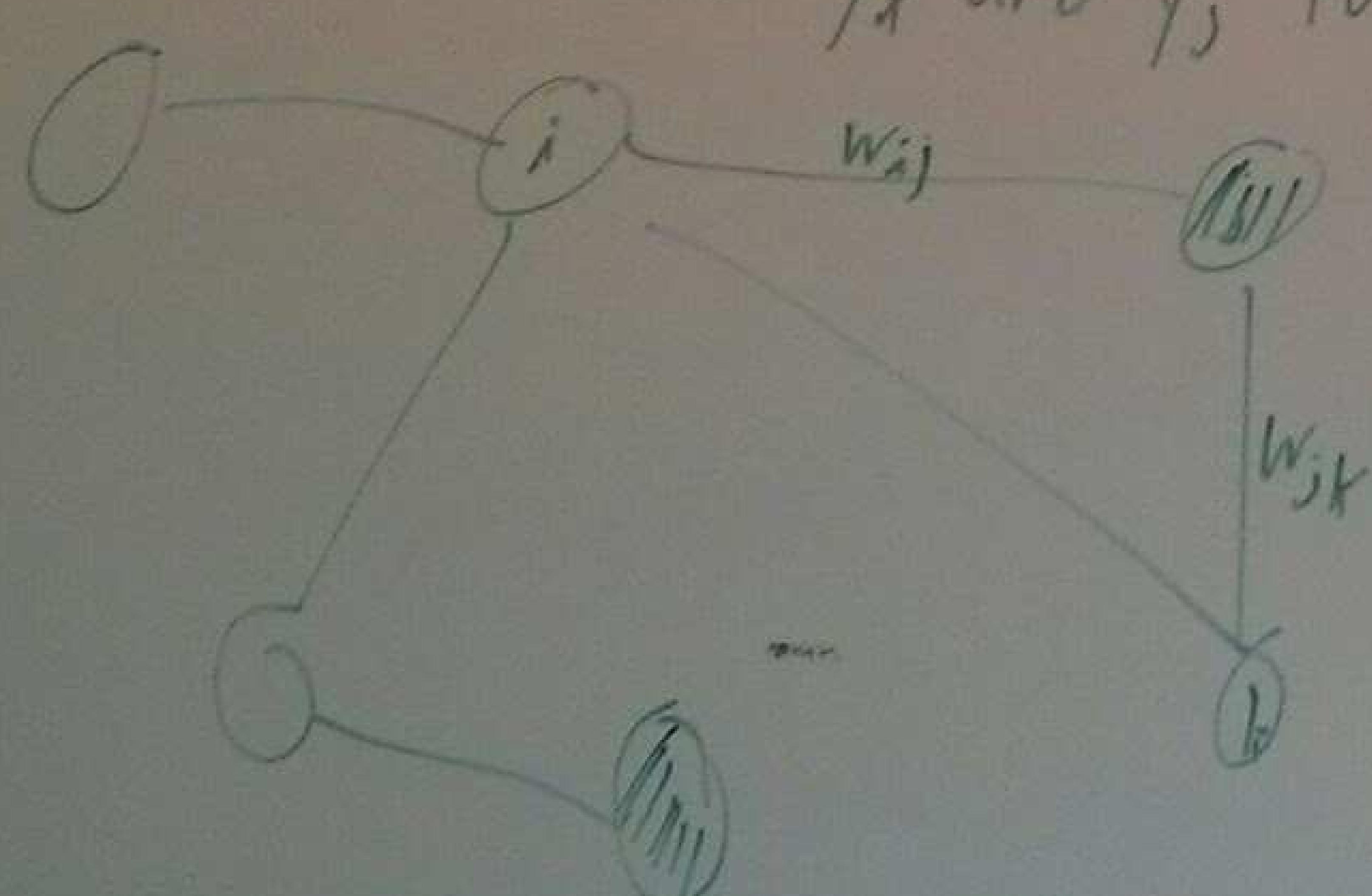
   - penalize entropy of $p(y_u | X_u, w)$

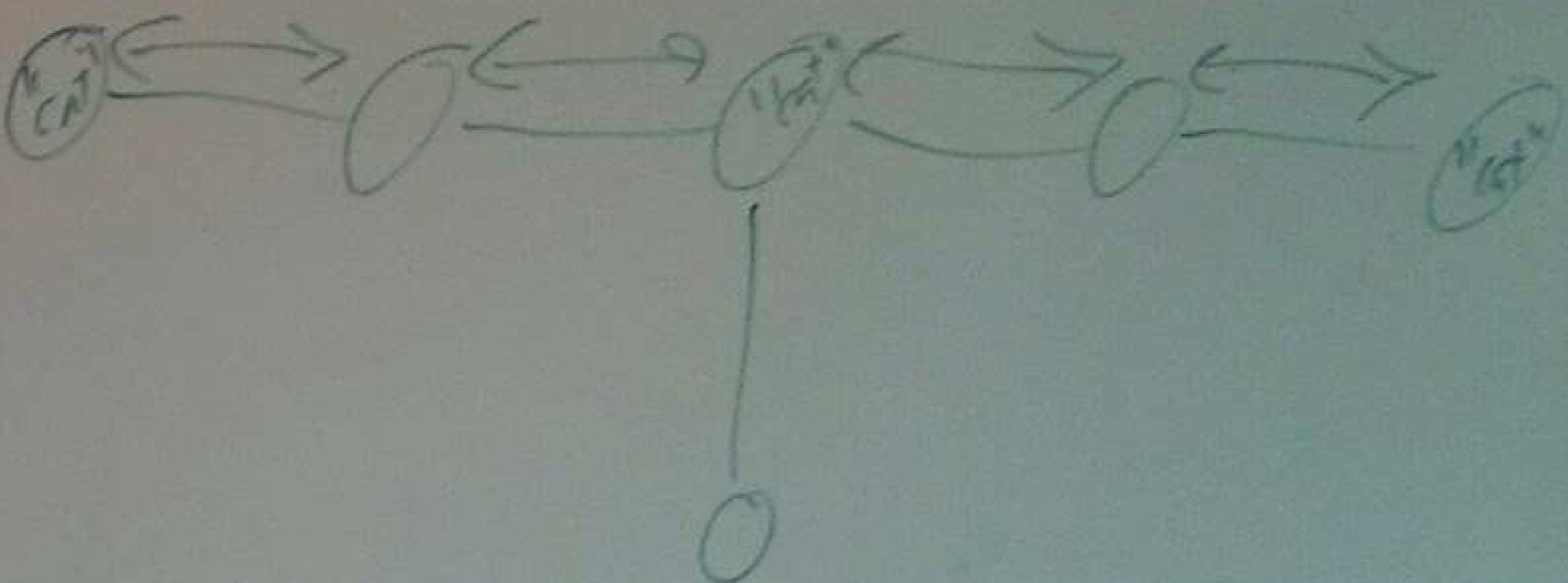
   - want labels to be non-random.

4. Transductive SVM: "just say no"

5. Graph-based SSL

- use features or relationships between x_i to define a graph.
- graph has a weight $w_{ij} \geq$ how much we want y_i and y_j to agree.





As

Problems with MAP estimation

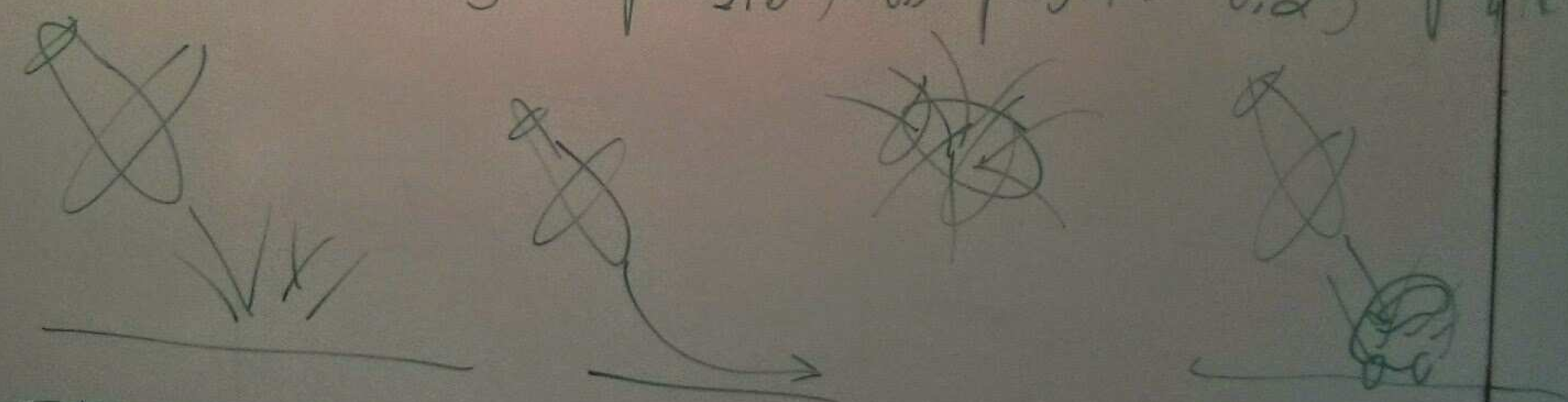
h: hypothesis
D: data
H: hypothesis space

- Does MAP make the right decision?

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

h_1 h_2 h_3 h_4

$p(h_1|D) = 0.25$ $p(h_2|D) = 0.3$ $p(h_3|D) = 0.25$ $p(h_4|D$



- Missing notion of risk
- $p(h_2|D) = 0.3$ (MAP) $p(\neg h_2|D) = 0.7$
- $p(\neg live|D) = p(h_1|D) + p(h_3|D) + p(h_4|D) = 0.7$

Learning principles

ML: $\hat{h} = \underset{h}{\operatorname{argmax}}\ p(D|h)$

predict using $\underset{\hat{D}}{\operatorname{argmax}}\ p(\hat{D}|\hat{h})$

MAP $\hat{h} = \underset{h}{\operatorname{argmax}}\ p(h|D) \propto p(D|h)p(h)$

predict using $\underset{\hat{D}}{\operatorname{argmax}}\ p(\hat{D}|\hat{h})$

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

(not a "point" estimate)

predict by integrating over "hidden" parameters

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

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

hidden: $\hat{h} = \underset{h}{\operatorname{argmax}} \sum_z p(h,z|D)$

bagging

$$\hat{h} = \sum_{D_i \in \text{bootstrap}(D)} \underset{h}{\operatorname{argmax}}\ p(h|D)$$

yesian learning

njugate priors

Example: Coin flipping

Bernoulli likelihood

$$p(X='H'|\theta)=\theta$$

$$p(X='T'|\theta)=(1-\theta)$$

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

Beta prior on θ

Uniform: Beta(1,1)

$$\theta \sim Beta(a,b)$$

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

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

$$\int p(\theta|a,b)d\theta=1$$

$$\int \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}d\theta=1$$

$$\int \theta^{a-1}(1-\theta)^{b-1}d\theta=B(a,b)$$

Posterior

- Assume we observe 'HHH'

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

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

$$= \theta^3 \theta^{a-1} (1-\theta)^{b-1} / p(HHH)$$

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

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

(hid $p(HHH \mid a, b)$)

$= B(3+a, b)$

"marginal likelihood"

...

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

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

mean of posterior:

$$\frac{\alpha}{\alpha + b} = \frac{(3+a)}{(3+a)+b} \underset{\text{uniform}}{=} \frac{4}{5}$$

$\approx 80\%$ heads.

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

$a=.1, b=.1$

$a=100, b=1$

$$p(\hat{H} \mid HHH)$$

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

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

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

$$= \frac{3+a}{3+a+b}$$

(not ya