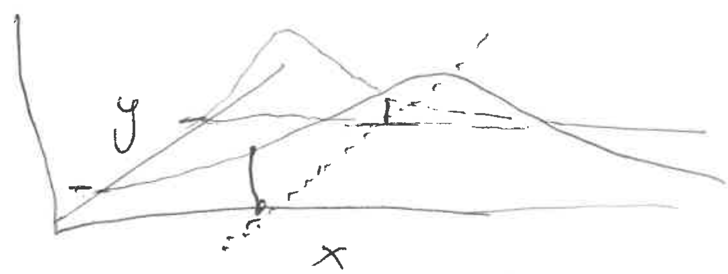
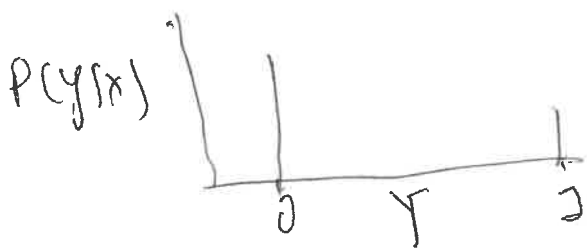
I Bayes Classification.

$P(X, Y) =$



$P(y|x) =$



(1) Pick the most likely one if you dont want to be wrong.

If you made N decisions, avg you'd be wrong on $N(P(\bar{y}|x))$ times.

II But now assign costs.

- $C(\text{cold} \to \text{cold}) \to$ Take meds.
- $C(\text{cold} \to Nc) \to$ Admitted to ER
- $C(Nc \to \text{cold}) \to$ Poison yourself with meds
- $C(Nc \to Nc) \to$ Nothing.

$$E(C(\text{cold})) = \text{Cost}(\text{cold} \to \text{cold}) \cdot P(\text{Cold} | x) + \text{Cost}(Nc \to \text{cold}) \cdot P(Nc | x)$$

$$E(C(Nc)) = \text{Cost}(\text{cold} \to Nc)\ P(\text{Cold} | x) + \text{cost}(Nc \to Nc)\ P(Nc | x).$$

find the min

Minimum Bayes Risk classification.

$$\hat{Y} = \underset{Y}{\arg\min}\ E(c(Y))$$

$$= \underset{Y}{\arg\min} \sum_{y^*} c(y^* \to y)\, P(Y^* | x)$$

$$\underset{Y}{\arg\min} \sum_{Y^*} Cost(y^*; y)\, P(Y^* | x).$$

For cost : 1/0 (0 for $y^* = y$, 1 else)

$$\hat{Y} = \underset{Y}{\arg\min} \sum_{Y^* \neq Y} P(Y^* | x) = \underset{Y^* \neq Y}{\arg\min} (1 - P(Y|x))$$

$$= \underset{Y}{\arg\max}\ P(Y|x) \longrightarrow$$ Usual MAP classifier

The above example was simple. Extend to Structured Pred setup.

$$\hat{Y} = \underset{Y}{\arg\min}\ E(Cost(Y))$$

$$= \underset{Y}{\arg\min} \sum_{Y^*} Cost(Y^*, Y)\, P(Y^* | x)$$

$\hookrightarrow$ Exponential sum.

(i) Simple example. Y = state seq.

$P(X, Y)$ = HMM.

Cost = 1/0

soln = ? $\quad \underset{Y}{\arg\max}\ P(Y|x)$

Viterbi

2 Simple example -

$P(X,Y) =$ PCFG. $\quad Y =$ tree

Cost = 1/0

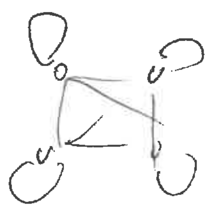
Soln = ? $\quad \arg\max_{Y} P(Y|X)$

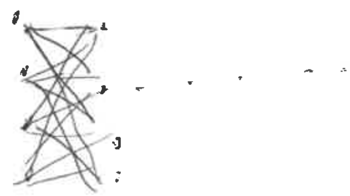
CYK

More generally, MBR decoding!

Go to P 5

HMM variant





MAP decoding is the ~~first~~ MBR with 0/1 cost.

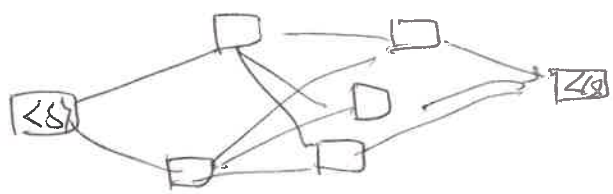
TRIVIAL solution 1.

Not defining $C(Y^*, Y)$, keeping it generic within our problem setting.

Problem → Define a loss between 2 word strings. $L(w^*, w)$.

E.g. Levenshtein distance, or distance which makes some kind of errors more important

Basic Setup :⇒ A word graph.



Nodes & Edges have log probs.

Algo 1: Hack, greedy.

Use A^* to find N most likely decodes.

$\left.\begin{matrix} W_1^* \\ \vdots \\ W_N^* \end{matrix}\right\}$ Candidates.

$\hat{w} = \underset{i}{\mathrm{argmin}} \; E\left[\mathrm{Cost}(W_i^*)\right]$

How do you compute this cost?

Posterior decoding with constraints

$$\delta^{PK} = \underset{S \in P_A}{\arg\max} \prod_{i=1}^{L} P(S_i | O)$$ I can do this bcoz of synchronous decoding

subject to constraint $\underline{\delta(l_1, l_2)}$ (seq constr).

Init:

$V_{start}(0) = 1 \qquad V_k(0) = 0 \quad k \neq start$

Recurr

$$V_k(i) = \max_{\{s\}} \left(V_s(i-1)\, \delta'(s,k) \right) \frac{P(S_t = k | O)}{\gamma_t(k|O)}$$

$$P_i(k) = \arg\max_{\{s\}} [\quad]$$

Terminati

$$P(\delta^{Pk} | O) = \max_s V_s[L]\, \delta'(s, END)$$

$$S_L^{Pk} = \arg\max_{\{s\}}$$

Track.

$$S_{t-1}^{PC} = P_t(S_t^{PC}).$$

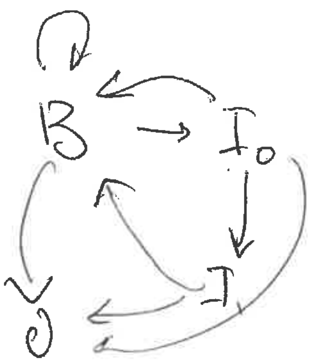
Assignment: $A_i = \text{label}(S_{ET}^{PC})$

[Draw diagram]

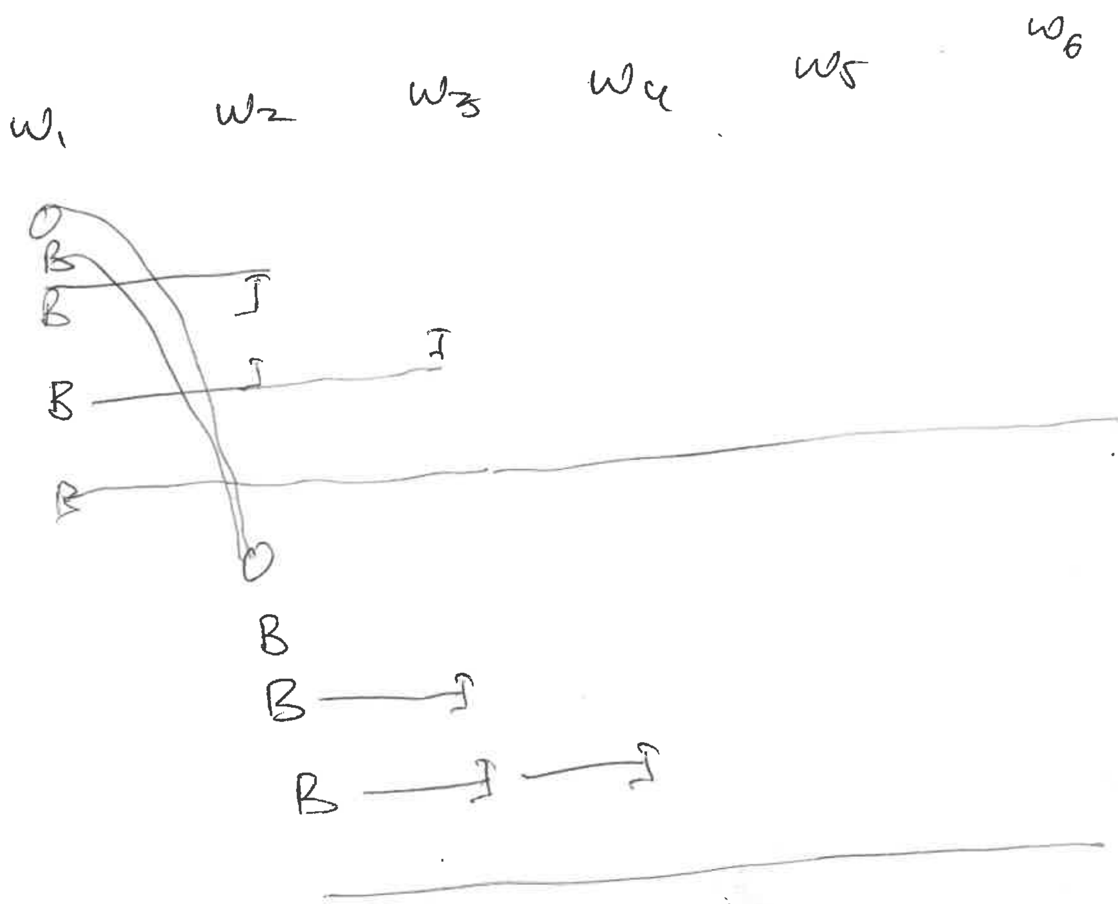
If we have multiple states for the same label, ~~average~~ sum first

How about when you have constraints?

E.g BIO tagger

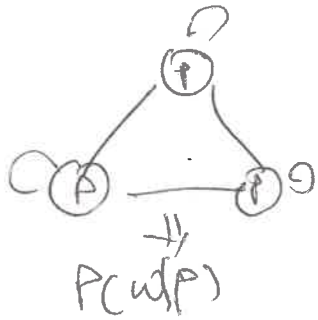


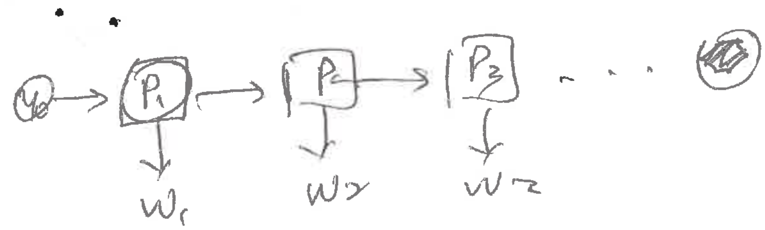
Now you end up with something that looks like Viterbi but has to keep track of sequences too



Constrained posterior decoding.

HMM for POS tagging





Cost = # mislabelled tokens.

$$= \sum_{i=1}^{n} \mathbb{I}(\text{label}(i) \neq \text{label}^*(i))$$

$$= \sum_{i=1}^{n} \mathbb{I}(y_i \neq y_i^*)$$

Expected cost $= E \sum_{i=1}^{n} \mathbb{I}(y_i \neq y_i^*)$

$$\underset{Y}{\text{argmin}} \; \text{Cost} = \underset{y_1 \dots y_N}{\text{argmin}} \; E \sum_{i=1}^{n} \mathbb{I}(\;)$$

$$= \underset{y_1 \dots y_N}{\text{argmin}} \sum_{i=1}^{n} E\, \mathbb{I}(y_i \neq y_i^*)$$

$$\Rightarrow y_i = \underset{y}{\text{argmin}} \; E\, \mathbb{I}(y \neq y_i^*)$$

$$= \underset{y}{\text{argmin}} \sum_{y^*} P(y^*|x)\, \mathbb{I}(y = y_i^*)$$

$$= \underset{y}{\text{argmax}} \; P(y_i = y | x)$$

Its just a series of local decisions

$$T_G = \operatorname*{argmax}_T \; E(L/N_c) = \cancel{\tfrac{1}{N_c}} \operatorname*{argmax}_T \cancel{E}\, \mathbb{I}((s,t,R) \in T)$$

$$= \underline{\operatorname*{argmax}_T \sum_N P(s,t,R \mid w).}$$

$$T_G = \operatorname*{argmax}_T \sum_{T_c} P(T_c \mid w_1^n)\, |T \cap T_c|$$

$$= \operatorname*{argmax}_T \sum_{T_c} P(T_c \mid w_1^n) \sum_{(s,t,x) \in T} \mathbb{I}((s,t,x) \in T_c)$$

For a PCFG.

$$P(S \to w_1^{s-1}\, X\, w_{t+1}^n \mid w_1^n) = \sum_{T_c} P(T_c \mid w_1^n)\, \mathbb{I}((s,t,x) \in T_c)$$

$$T_G = \operatorname*{argmax}_T \sum_{(s,t,x) \in T} P(S \to w_1^{s-1}\, X\, w_{t+1}^n \mid w_1^n)$$

But $P(S \to w_1^{s-1}\, X\, w_{t+1}^n \mid w_1^n)$

$$= \frac{P(S \to \cdots, w_1^n)}{P(S \to w_1^n)} = \frac{P(S \to \cdots X \cdots)\, P(X \to w_s^t)}{P(S \to w_1^n)}$$

$$= \frac{\beta(s,t)\, \alpha(s,t)}{P(S \to w_1^n)}$$

$$\Rightarrow T_G = \operatorname*{argmax}_T \sum_{(s,t,x) \in T} \gamma(s,t,x)$$

Goodman '96

labeled match $(s,t,R) \to (s,t,R)$

bracketed match $(s,t,R) \to (s,t,*)$

CYK etc optimize labeled TREE match.

$T^* = T$, $L/N_c = 1 \Rightarrow$ cost = 0, else = 1

L = no of labelled brackets matching.

V. Imp for eg Travel agent

- "find me all flights on tuesday"

If we split it wrong, we'll wait till tuesday & get a wrong ans.

- But in MT "his credentials are nothing which should be ~~matche~~ laughed at!

& MT mis aligns → 'His creds are nothing, which should make you laugh' ≠ good,

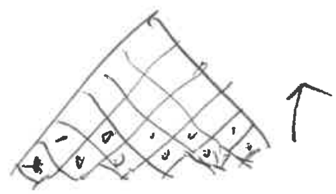
But still helps.

Here we want labelled RECALL.

$$T_c = \underset{T_t}{\operatorname{argmax}}\ L/N_c$$

Algo for max recall rate parse.

$$\text{MaxC}(s,t) = \max_{x} g(s,t,x) + \max_{r \mid s \le r < t} \left(\text{MaxC}(s,r) + \text{MaxC}(r+1,t) \right)$$



for r rules, k NTs, $O(n^3 + kn^2)$

Dominated by outside prob computation.