Hidden Markov Models

He He

New York University

October 11, 2020

Generative vs discriminative models

Generative modeling: $p(x, y)$

Discriminative modeling: $p(y \mid x)$

Examples:

	generative	discriminative
classification	Naive Bayes	logistic regression
sequence labeling	HMM	CRF

Table of Contents

Generative modeling for sequence labeling

DT	NN	VBD	IN	DT	NN
$\uparrow$	$\uparrow$	$\uparrow$	$\uparrow$	$\uparrow$	$\uparrow$
the	fox	jumped	over	the	dog

Task: given $x = (x_1, \ldots, x_m) \in \mathcal{X}^m$, predict $y = (y_1, \ldots, y_m) \in \mathcal{Y}^m$

Three questions:

- Modeling: how to define a parametric joint distribution $p(x, y; \theta)$?
- Learning: how to estimate the parameters θ given observed data?
- Inference: how to efficiently find $\arg\max_{y \in \mathcal{Y}^m} p(x, y; \theta)$ given x?

Decompose the joint probability

$$\begin{tikzcd}
\mathrm{DT} \arrow[r] \arrow[d] & \mathrm{NN} \arrow[r] \arrow[d] & \mathrm{VBD} \arrow[r] \arrow[d] & \mathrm{IN} \arrow[r] \arrow[d] & \mathrm{DT} \arrow[r] \arrow[d] & \mathrm{NN} \arrow[d] \\
\text{the} & \text{fox} & \text{jumped} & \text{over} & \text{the} & \text{dog}
\end{tikzcd}$$

$$\begin{aligned}
p(x, y) &= p(x \mid y)p(y) \\
&= p(x_1, \ldots, x_m \mid y)p(y) \\
&= \prod_{i=1}^{m} p(x_i \mid y)p(y) \quad \text{Naive Bayes assumption} \\
&= \prod_{i=1}^{m} p(x_i \mid y_i)p(y_1, \ldots, y_m) \quad \text{a word only depends its own tag} \\
&= \prod_{i=1}^{m} p(x_i \mid y_i) \prod_{i=1}^{m} p(y_i \mid y_{i-1}) \quad \text{Markov assumption}
\end{aligned}$$

Hidden Markov models

Hidden Markov model (HMM):

- Discrete-time, discrete-state Markov chain
- Hidden states $z_i \in \mathcal{Y}$ (e.g. POS tags)
- Observations $x_i \in \mathcal{X}$ (e.g. words)

$$p(x_{1:m}, y_{1:m}) = \prod_{i=1}^{m} \underbrace{p(x_i \mid y_i)}_{\text{emission probability}} \prod_{i=1}^{m} \underbrace{p(y_i \mid y_{i-1})}_{\text{transition probability}}$$

For sequence labeling:

- Transition probabilities: $p(y_i = t \mid y_{i-1} = t') = \theta_{t|t'}$ $O(|Y|^2 + 2|Y|)$
- Emission probabilities: $p(x_i = w \mid y_i = t) = \gamma_{w|t}$ $O(|X||Y|)$
- $y_0 = *, y_m = \texttt{STOP}$

Learning: MLE

Data: $\mathcal{D} = \{(x, y)\}$ $(x \in \mathcal{X}^m, y \in \mathcal{Y}^m)$
Task: estimate transition probabilities $\theta_{t|t'}$ and emission probabilities $\gamma_{w|t}$ (# parameters?)

$$\ell(\theta, \gamma) = \sum_{(x,y)\in\mathcal{D}} \left(\sum_{i=1}^{m} \log p(x_i \mid y_i) + \sum_{i=1}^{m} \log p(y_i \mid y_{i-1}) \right)$$

$$\max_{\theta,\gamma} \sum_{(x,y)\in\mathcal{D}} \left(\sum_{i=1}^{m} \log \gamma_{x_i|y_i} + \sum_{i=1}^{m} \log \theta_{y_i|y_{i-1}} \right)$$

$$\text{s.t.} \quad \sum_{w\in\mathcal{X}} \gamma_{w|t} = 1 \quad \forall w \in \mathcal{X}$$

$$\sum_{t\in\mathcal{Y}\cup\{\texttt{STOP}\}} \theta_{t|t'} = 1 \quad \forall t' \in \mathcal{Y} \cup \{*\}$$

MLE solution

Count the occurrence of certain transitions and emissions in the data.

Transition probabilities:

$$\theta_{t|t'} = \frac{\text{count}(\overset{\text{PT}}{t'} \rightarrow \overset{\text{NN}}{t})}{\sum_{a \in \mathcal{Y} \cup \{\texttt{STOP}\}} \text{count}(\underset{\text{PT}}{t'} \rightarrow a)}$$

Emission probabilities:

$$\gamma_{w|t} = \frac{\text{count}(w, t)}{\sum_{w' \in \mathcal{X}} \text{count}(w', t)}$$

Inference

Task: given $x \in \mathcal{X}^m$, find the most likely $y \in \mathcal{Y}^m$

$$
\begin{aligned}
&\arg\max_{y \in \mathcal{Y}^m} \log p(x, y) \\
&= \arg\max_{y \in \mathcal{Y}^m} \sum_{i=1}^{m} \log p(x_i \mid y_i) + \sum_{i=1}^{m} \log p(y_i \mid y_{i-1})
\end{aligned}
$$

Viterbi + backtracking:

$$
\pi[j, t] = \max_{t' \in \mathcal{Y}} \left(\log p(x_j \mid t) + \log p(t \mid t') + \pi[j-1, t']\right)
$$

Table of Contents

Naive Bayes with missing labels

Task:

- Assume data is generated from a Naive Bayes model.
- Observe $\left\{x^{(i)}\right\}_{i=1}^{N}$ without labels.
- Estimate model parameters and the most likely labels.

ID	US	government	gene	lab	label
1	1	1	0	0	?
2	0	1	0	0	?
3	0	0	1	1	?
4	0	1	1	1	?
5	1	1	0	0	?

A chicken and egg problem

If we know the model parameters, we can predict labels easily.
If we know the labels, we can estiamte the model parameters easily.

Idea: start with guesses of labels, then iteratively refine it.

ID	US	government	gene	lab	label
1	1	1	0	0	
2	0	1	0	0	
3	0	0	1	1	
4	0	1	1	1	
5	1	1	0	0	

	US	government	gene	lab
$p(\cdot \mid 0)$				
$p(\cdot \mid 1)$				

A chicken and egg problem

If we know the model parameters, we can predict labels easily.
If we know the labels, we can estiamte the model parameters easily.

Idea: start with guesses of labels, then iteratively refine it.

$\arg\max_y p(y|x)$

$p(y=1|x)$

$\propto p(x|y)p(y)$

$= \frac{1}{2} \times 1 \times \frac{2}{5} = \frac{1}{5} =$

$p(y=0|x)$

$\propto \frac{1}{3} \times \frac{2}{3} \times \frac{3}{5}$

$= \frac{2}{15}$

ID	US	government	gene	lab	label
1	1	1	0	0	0
2	0	1	0	0	0
3	0	0	1	1	0
4	0	1	1	1	1
5	1	1	0	0	1

1	0
$\frac{1}{5}$	$\frac{2}{15}$

	US	government	gene	lab
$p(\cdot \mid 0)$	$1/3$	$2/3$	$1/3$	$1/3$
$p(\cdot \mid 1)$	$1/2$	1	$1/2$	$1/2$

Algorithm: EM for NB

1. Initialization: $\theta \leftarrow$ random parameters
2. Repeat until convergence:
   (i) Inference:
   $$q(y \mid x^{(i)}) = p(y \mid x^{(i)}; \theta)$$
   (ii) Update parameters:
   $$\theta_{w|y} = \frac{\sum_{i=1}^{N} q(y \mid x^{(i)}) \mathbb{I}\left[w \text{ in } x^{i}\right]}{\sum_{i=1}^{N} q(y \mid x^{(i)})}$$

- With fully observed data, $q(y \mid x^{(i)}) = 1$ if $y^{(i)} = y$.
- Similar to the MLE solution except that we're using "soft counts".
- What is the algorithm optimizing?

Objective: maximize marginal likelihood

Likelihood: $L(\theta; \mathcal{D}) = \prod_{x \in \mathcal{D}} p(x; \theta)$

Marginal likelihood: $L(\theta; \mathcal{D}) = \prod_{x \in \mathcal{D}} \overbrace{\sum_{z \in \mathcal{Z}} p(x, z; \theta)}^{\text{marginal prob}}$

- Marginalize over the (discrete) latent variable $z \in \mathcal{Z}$ (e.g. missing labels)

Maximum marginal log-likelihood estimator:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \sum_{x \in \mathcal{D}} \underbrace{\log \sum_{z \in \mathcal{Z}} p(x, z; \theta)}_{\log P(x)}$$

Goal: maximize $\log p(x; \theta)$
Challenge: in general not concave, hard to optimize

Intuition

Problem: marginal log-likelihood is hard to optimize (only observing the words)

Observation: complete data log-likelihood is easy to optimize (observing both words and tags)

$$\max_{\theta} \log p(x, z; \theta)$$

Idea: guess a distribution of the latent variables $q(z)$ (soft tags)

Maximize the expected complete data log-likelihood:

$$\max_{\theta} \sum_{z \in \mathcal{Z}} q(z) \log p(x, z; \theta)$$

EM assumption: the expected complete data log-likelihood is easy to optimize (use soft counts)

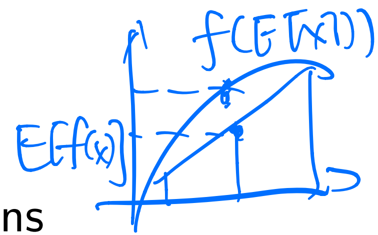
Lower bound of the marginal log-likelihood

$f(z)$

$$
\begin{aligned}
\log p(x;\theta) &= \log \sum_{z\in\mathcal{Z}} p(x,z;\theta) \\
&= \log \underbrace{\sum_{z\in\mathcal{Z}} q(z)}\frac{p(x,z;\theta)}{q(z)} \qquad = \log \mathbb{E}_z\left[\frac{p(x,z;\theta)}{q(z)}\right] \\
&\geq \sum_{z\in\mathcal{Z}} q(z)\log\frac{p(x,z;\theta)}{q(z)} \qquad = \mathbb{E}_z\left[\log \frac{p(x,z;\theta)}{q(z)}\right] \\
&\stackrel{\text{def}}{=} \mathcal{L}(q,\theta)
\end{aligned}
$$

Jensen's inequality

$E[f(X)] \leq f(E[X])$



- **Evidence**: $\log p(x;\theta)$
- **Evidence lower bound (ELBO)**: $\mathcal{L}(q,\theta)$
- q: chosen to be a family of tractable distributions
- Idea: maximize the ELBO instead of $\log p(x;\theta)$

Justification for maximizing ELBO

$$\mathrm{KL}(p \| q) = E_p\left[\log \frac{p(x)}{q(x)}\right]$$

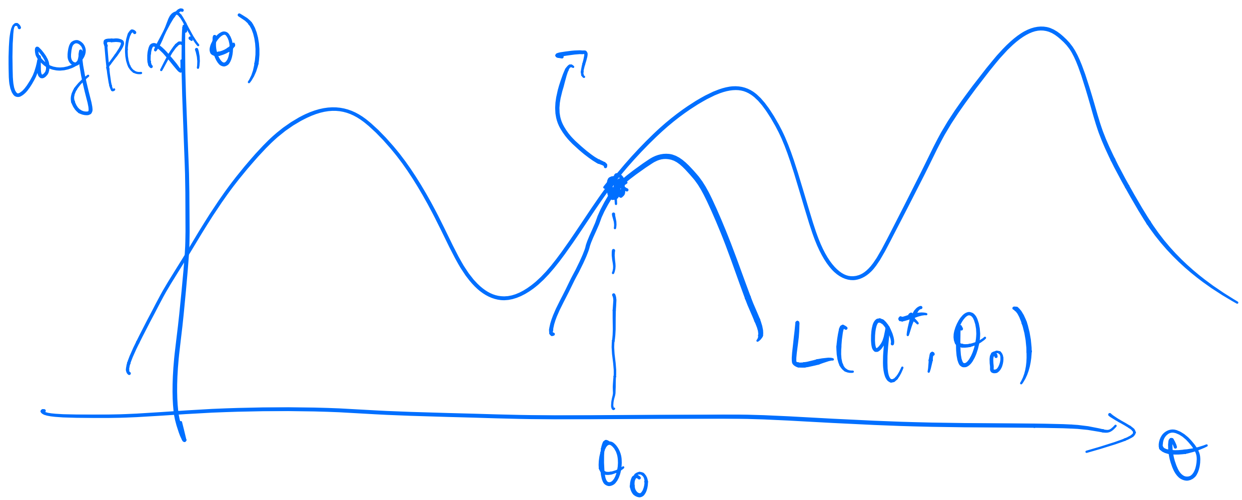
$$
\begin{aligned}
\mathcal{L}(q,\theta) &\overset{\text{def}}{=} \sum_{z\in\mathcal{Z}} q(z)\log\frac{p(x,z;\theta)}{q(z)} \\
&= \sum_{z\in\mathcal{Z}} q(z)\log\frac{p(z\mid x;\theta)p(x;\theta)}{q(z)} \\
&= -\sum_{z\in\mathcal{Z}} q(z)\log\frac{q(z)}{p(z\mid x;\theta)} + \sum_{z\in\mathcal{Z}} q(z)\log p(x;\theta) \\
&= \underset{\leq 0}{-\mathrm{KL}\left(q(z)\|p(z\mid x;\theta)\right)} + \underbrace{\log p(x;\theta)}_{\text{evidence}}
\end{aligned}
$$

- **KL divergence**: measures "distance" between two distributions (not symmetric!) $\mathrm{KL}(p\|q) \neq \mathrm{KL}(q\|p)$
- $\mathrm{KL}(q\|p) \geq 0$ with equality iff $q(z) = p(z \mid x)$.
- ELBO $=$ evidence - KL $\leq$ evidence

Justification for maximizing ELBO

$$\mathcal{L}(q,\theta) = -\overset{=0}{\mathrm{KL}\left(q(z) \| p(z \mid x;\theta)\right)} + \log p(x;\theta)$$

Fix $\theta = \theta_0$ and $\max_q \mathcal{L}(q,\theta_0)$: $q^* = p(z \mid x;\theta_0)$



Let θ^*, q^* be the global optimzer of $\mathcal{L}(q,\theta)$, then θ^* is the global optimizer of $\log p(x;\theta)$. (Proof: exercise)

Summary

Latent variable models: clustering, latent structure, missing lables etc.

Parameter estimation: maximum marginal log-likelihood

Challenge: directly maximize the **evidence** $\log p(x; \theta)$ is hard

Solution: maximize the **evidence lower bound**:

$$\text{ELBO} = \mathcal{L}(q, \theta) = -\text{KL}\left(q(z) \| p(z \mid x; \theta)\right) + \log p(x; \theta)$$

Why does it work?

$$\begin{aligned} q^*(z) &= p(z \mid x; \theta) \quad \forall \theta \in \Theta \\ \mathcal{L}(q^*, \theta^*) &= \max_\theta \log p(x; \theta) \end{aligned}$$

EM algorithm

"Coordinate ascent" on $\mathcal{L}(q, \theta)$

1. Random initialization: $\theta^{\text{old}} \leftarrow \theta_0$
2. Repeat until convergence

   (i) $q(z) \leftarrow \arg\max_q \mathcal{L}(q, \theta^{\text{old}})$

$$\textbf{Expectation} \text{ (the E-step):} \quad q^*(z) = p(z \mid x; \theta^{\text{old}})$$

$$\text{ELBO} = L(q^*, \theta) = J(\theta) = \sum_{z \in \mathcal{Z}} q^*(z) \log \frac{p(x, z; \theta)}{q^*(z)}$$

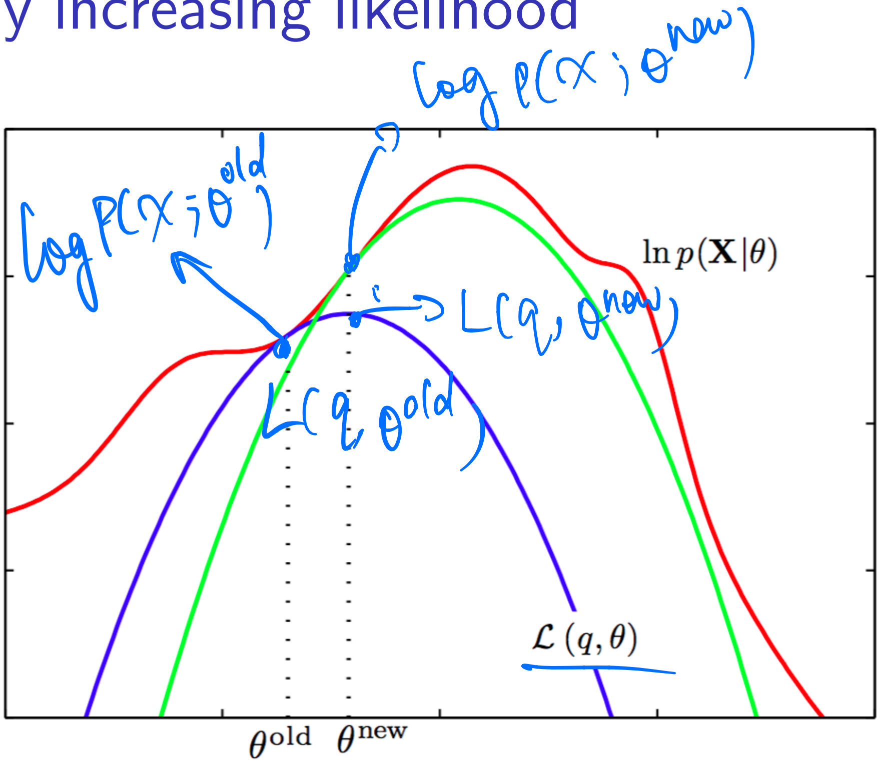
   (ii) $\theta^{\text{new}} \leftarrow \arg\max_\theta \mathcal{L}(q^*, \theta)$

Max ~~Mini~~mization (the M-step): $\theta^{\text{new}} \leftarrow \arg\max_\theta J(\theta)$

max expected complete data likelihood.

EM puts no constraint on q in the E-step and assumes the M-step is easy. In general, both steps can be hard.

Monotonically increasing likelihood



Exercise: prove that EM increases the marginal likelihood monotonically

$$\log p(x;\theta^{\text{new}}) \geq \log p(x;\theta^{\text{old}}) \; .$$

Does EM converge to a global maximum?

EM for multinomial naive Bayes

Setting: $x = (x_1, \ldots, x_m) \in \mathcal{V}^m, z \in \{1, \ldots, K\}, \mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$

E-step:

$$q^*(z) = p(z \mid x; \theta^{\text{old}}) = \frac{\prod_{i=1}^m p(x_i \mid z; \theta^{\text{old}}) p(z; \theta^{\text{old}})}{\sum_{z' \in \mathcal{Z}} \prod_{i=1}^m p(x_i \mid z'; \theta^{\text{old}}) p(z'; \theta^{\text{old}})}$$

$$J(\theta) = \sum_{x \in \mathcal{D}} \sum_{z \in \mathcal{Z}} q_x^*(z) \log p(x, z; \theta) = \sum_{x \in \mathcal{D}} \sum_{z \in \mathcal{Z}} q_x^*(z) \log \underbrace{\prod_{i=1}^m p(x_i \mid z; \theta) p(z; \theta)}_{\text{NB}}$$

the fox jumped.

M-step:

$\max_\theta J(\theta)$

$$\max_\theta \sum_{x \in \mathcal{D}} \sum_{z \in \mathcal{Z}} q_x^*(z) \left(\sum_{w \in \mathcal{V}} \log \theta_{w|z}^{\text{count}(w|x)} + \log \theta_z \right)$$

$$\text{s.t.} \quad \sum_{w \in \mathcal{V}} \theta_{w|z} = 1 \quad \forall w \in \mathcal{V}, \quad \sum_{z \in \mathcal{Z}} \theta_z = 1,$$

where $\text{count}(w \mid x) \stackrel{\text{def}}{=} \#$ occurrence of w in x

EM for multinomial naive Bayes

M-step has closed-form solution:

$$\theta_z = \frac{\sum_{x\in\mathcal{D}} q_x^*(z)}{\sum_{z\in\mathcal{Z}} \sum_{x\in\mathcal{D}} \underbrace{q_x^*(z)}_{\text{soft label count}}}$$

$$\theta_{w|z} = \frac{\sum_{x\in\mathcal{D}} q_x^*(z)\text{count}(w \mid x)}{\sum_{w\in\mathcal{V}} \sum_{x\in\mathcal{D}} \underbrace{q_x^*(z)\text{count}(w \mid x)}_{\text{soft word count}}}$$

Similar to the MLE solution except that we're using soft counts.

M-step for multinomial naive Bayes

$$\max_{\theta} \sum_{x \in \mathcal{D}} \sum_{z \in \mathcal{Z}} q_x^*(z) \left(\sum_{w \in \mathcal{V}} \log \theta_{w|z}^{\text{count}(w|x)} + \log \theta_z \right)$$

$$\text{s.t.} \quad \sum_{w \in \mathcal{V}} \theta_{w|z} = 1 \quad \forall w \in \mathcal{V}, \quad \sum_{z \in \mathcal{Z}} \theta_z = 1$$

Summary

Expectation max **~~mini~~mization (EM)** algorithm: maximizing ELBO $\mathcal{L}(q, \theta)$ by coordinate ascent

E-step: Compute the expected complete data log-likelihood $J(\theta)$ using $q^*(z) = p(z \mid x; \theta^{\text{old}})$

M-step: Maximize $J(\theta)$ to obtain θ^{new}

Assumptions: E-step and M-step are easy to compute

Properties: Monotonically improve the likelihood and converge to a stationary point

Table of Contents

HMM recap

$P(x|y)P(y)$

$\prod_{i=1}^{m} P(x_i|y)$

$\prod_{i=1}^{m} P(x_i|y_i)$

Setting:

- Hidden states $z_i \in \mathcal{Y}$ (e.g. POS tags)
- Observations $x_i \in \mathcal{X}$ (e.g. words)

$$p(x_{1:m}, y_{1:m}) = \prod_{i=1}^{m} \underbrace{p(x_i \mid y_i)}_{\text{emission probability}} \prod_{i=1}^{m} \underbrace{p(y_i \mid y_{i-1})}_{\text{transition probability}}$$

Parameters:

- Transition probabilities: $p(y_i = t \mid y_{i-1} = t') = \theta_{t|t'}$
- Emission probabilities: $p(x_i = w \mid y_i = t) = \gamma_{w|t}$
- $y_0 = *, y_m = \texttt{STOP}$

Task: estimate parameters given incomplete observations

E-step for HMM

E-step:

$$
\begin{aligned}
q^*(z) &= p(z \mid x; \theta, \gamma) \\
\mathcal{L}(q^*, \theta, \gamma) &= \sum_{x \in \mathcal{D}} \underbrace{\sum_{z \in \mathcal{Z}} q_x^*(z) \log p(x, z; \theta, \gamma)}_{\text{expected complete log-likelihood}} \\
&= \sum_{x \in \mathcal{D}} \sum_{z \in \mathcal{Z}} q_x^*(z) \log \underbrace{\prod_{i=1}^{m} p(x_i \mid z_i) p(z_i \mid z_{i-1})}_{\text{HMM}} \\
&= \sum_{x \in \mathcal{D}} \sum_{z \in \mathcal{Z}} q_x^*(z) \sum_{i=1}^{m} \left(\log \underbrace{p(x_i \mid z_i; \gamma)}_{\gamma_{x_i \mid z_i}} + \log \underbrace{p(z_i \mid z_{i-1}; \theta)}_{\theta_{z_i \mid z_{i-1}}} \right)
\end{aligned}
$$

M-step for HMM

M-step (similar to the NB solution):

$$\max_{\theta,\gamma} \mathcal{L}(q^*, \theta, \gamma) = \sum_{x\in\mathcal{D}} \sum_{z\in\mathcal{Z}} q_x^*(z) \sum_{i=1}^{m} \left(\log \gamma_{x_i|z_i} + \log \theta_{z_i|z_{i-1}}\right)$$

Emission probabilities:

$$\gamma_{w|t} = \frac{\sum_{x\in\mathcal{D}} \sum_{z\in\mathcal{Z}} q_x^*(z) \text{count}(w, t \mid x, z)}{\sum_{w'\in\mathcal{X}} \sum_{x\in\mathcal{D}} \sum_{z\in\mathcal{Z}} q_x^*(z) \text{count}(w', t \mid x, z)}$$

$$\text{count}(w, t \mid x, z) \stackrel{\text{def}}{=} \#\ \text{word-tag pairs}\ (w, t)\ \text{in}\ (x, z)$$

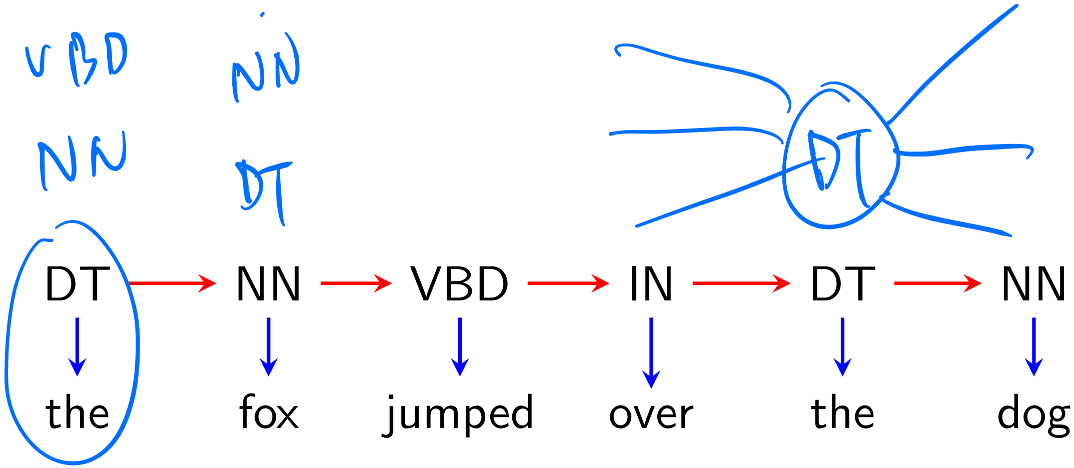
Transition probabilities:

$$\theta_{t|t'} = \frac{\sum_{x\in\mathcal{D}} \sum_{z\in\mathcal{Z}} q_x^*(z) \text{count}(t' \to t \mid z)}{\sum_{a\in\mathcal{Y}} \sum_{x\in\mathcal{D}} \sum_{z\in\mathcal{Z}} q_x^*(z) \text{count}(t' \to a \mid z)}$$

$$\text{count}(t' \to t \mid z) \stackrel{\text{def}}{=} \#\ \text{tag bigrams}\ (t', t)\ \text{in}\ z$$

M-step for HMM

Challenge: $\sum_{z \in \mathcal{Y}^m} q_x^*(z) \text{count}(w, t \mid x, z)$



Group sequences where $z_i = t$:

$$\sum_{z \in \mathcal{Y}^m} q_x^*(z) \text{count}(w, t \mid x, z) = \sum_{i=1}^{m} \mu_x(z_i = t) \mathbb{I}\left[x_i = w\right]$$

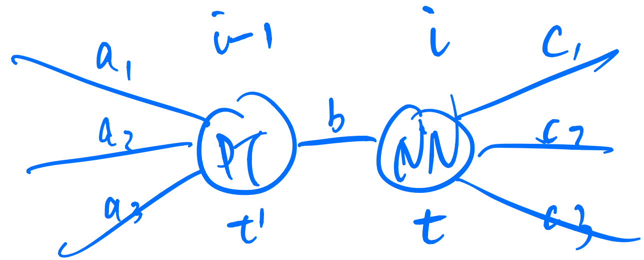
$$\mu_x(z_i = t) = \sum_{\{z \in \mathcal{Y}^m \mid z_i = t\}} q_x^*(z)$$

M-step for HMM

Challenge: $\sum_{z\in\mathcal{Y}^m} q_x^*(z)\text{count}(t' \to t \mid z)$

$a_1 b c_1 + a_1 b c_2 + \cdots$

$= (a_1 + a_2 + a_3) b (c_1 + c_2 + c_3)$



$$
\begin{array}{cccccc}
\text{DT} \rightarrow & \text{NN} \rightarrow & \text{VBD} \rightarrow & \text{IN} \rightarrow & \text{DT} \rightarrow & \text{NN} \\
\downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
\text{the} & \text{fox} & \text{jumped} & \text{over} & \text{the} & \text{dog}
\end{array}
$$

Group sequences where $z_i = t, z_{i-1} = t'$:

$$\sum_{z\in\mathcal{Y}^m} q_x^*(z)\text{count}(t' \to t \mid z) = \sum_{i=1}^{m} \mu_x(z_i = t, z_{i-1} = t')$$

$$\mu_x(z_i = t) = \sum_{\{z\in\mathcal{Y}^m \mid z_i = t, z_{i-1} = t'\}} q_x^*(z)$$

$z_{i-1} = t$

Compute tag marginals

$\mu_x(z_i = t)$: probability of the i-th tag being t given observed words x

$P(z|x)$ $\qquad P(z, x) / P(x)$

$$\mu_x(z_i = t) = \sum_{z:z_i=t} q_x^*(z) \propto \sum_{z:z_i=t} \prod_{j=1}^{m} \underbrace{q(x_i \mid z_i) q(z_i \mid z_{i-1})}_{\psi(z_i, z_{i-1})}$$

HMM

$$= \sum_{z:z_i=t} \prod_{j=1}^{i-1} \psi(z_j, z_{j-1}) \prod_{j=i}^{m} \psi(z_j, z_{j-1})$$

$z_{i+1:m}$

$$= \sum_{t'} \sum_{z:z_i=t, z_{i-1}=t'} \prod_{j=1}^{i-1} \psi(z_j, z_{j-1}) \prod_{j=i}^{m} \psi(z_j, z_{j-1})$$

$z_{1:i-2}$

$$= \sum_{t'} \left(\sum_{\substack{z_{1:i-1} \\ z_{i-1}=t'}} \prod_{j=1}^{i-1} \psi(z_j, z_{j-1}) \right) \psi(t, \cancel{t'}) \left(\sum_{\substack{z_{i+1:m} \\ z_i=t}} \prod_{j=i}^{m} \psi(z_j, z_{j-1}) \right)$$

$\psi(t' \to t)$ $\qquad j=i+1$

$$= \underbrace{\sum_{t'} \alpha[i-1, t] \psi(t, t')}\beta[i, t] = \alpha[i, t]\beta[i, t]$$

Compute tag marginals

Forward probabilities: probability of tag sequence prefix ending at $z_i = t$.

$$\alpha[i,t] \stackrel{\text{def}}{=} q(x_1,\ldots,x_i,z_i=t)$$
$$\alpha[i,t] = \sum_{t'\in\mathcal{Y}} \alpha[i-1,t']\psi(\cancel{t',t})$$

$\psi(t' \to t)$

Backward probabilities: probability of tag sequence suffix starting from z_{i+1} give $z_i = t$.

$$\beta[i,t] \stackrel{\text{def}}{=} q(x_{i+1},\ldots,x_m \mid z_i=t)$$
$$\beta[i,t] = \sum_{t'\in\mathcal{Y}} \beta[i+1,t']\psi(\cancel{t,t'})$$

$\psi(t \to t')$

Compute tag marginals

1. Compute forward and backward probabilities

$$\alpha[i,t] \quad \forall i \in \{1,\ldots,m\}, t \in \mathcal{Y} \cup \{\mathtt{STOP}\}$$
$$\beta[i,t] \quad \forall i \in \{m,\ldots,1\}, t \in \mathcal{Y} \cup \{*\}$$

2. Comptute the tag unigram and bigram marginals

$$\mu_x(z_i = t) \stackrel{\text{def}}{=} q(z_i = t \mid x)$$

$q(z_i, x)$

$$= \frac{\alpha[i,t]\beta[i,t]}{q(x)} = \frac{\alpha[i,t]\beta[i,t]}{\alpha[m,\mathtt{STOP}]}$$

$$\mu_x(z_{i-1} = t', z_i = t) \stackrel{\text{def}}{=} q(z_{i-1} = t', z_i = t \mid x)$$
$$= \frac{\alpha[i-1,t']\psi(t',t)\beta[i,t]}{q(x)}$$

In practice, compute in the log space.

Updated parameters

Emission probabilities:

$$\gamma_{w|t} = \frac{\sum_{x\in\mathcal{D}}\sum_{z\in\mathcal{Z}} q_x^*(z)\text{count}(w,t \mid x,z)}{\sum_{w'\in\mathcal{X}}\sum_{x\in\mathcal{D}}\sum_{z\in\mathcal{Z}} q_x^*(z)\text{count}(w',t \mid x,z)}$$
$$= \frac{\sum_{x\in\mathcal{D}}\sum_{i=1}^{m} \mu_x(z_i=t)\mathbb{I}\left[x_i=w\right]}{\sum_{w'\in\mathcal{X}}\sum_{x\in\mathcal{D}}\sum_{i=1}^{m} \mu_x(z_i=t)\mathbb{I}\left[x_i=w'\right]}$$

Transition probabilities:

$$\theta_{t|t'} = \frac{\sum_{x\in\mathcal{D}}\sum_{z\in\mathcal{Z}} q_x^*(z)\text{count}(t'\to t \mid z)}{\sum_{a\in\mathcal{Y}}\sum_{x\in\mathcal{D}}\sum_{z\in\mathcal{Z}} q_x^*(z)\text{count}(t'\to a \mid z)}$$
$$= \frac{\sum_{x\in\mathcal{D}}\sum_{i=1}^{m} \mu_x(z_{i-1}=t', z_i=t)}{\sum_{a\in\mathcal{Y}}\sum_{x\in\mathcal{D}}\sum_{i=1}^{m} \mu_x(z_{i-1}=t', z_i=a)}$$

Summary

EM for HMM:

1. Randomly initialize the emission and transition probabilities
2. Repeat until convergence
   (i) Compute forward and backward probabilities
   (ii) Update the emission and transition probabilities using expected counts

   If the solution is bad, re-run EM with a different random seed.

General EM:

- One example of variational methods (use a tractable q to approximate p)
- May need approximation in both the E-step and the M-step
- Useful in probabilistic models and Bayesian methods