Expectation Maximization

A Method for estimating the parameters of a probabilistic model
a.k.a.
An algorithm for Machine Learning

Slides prepared by Pushpendre Rastogi

Preliminaries

- PRML uses chapter 9.1-9.3 to gently introduce EM
  - I will start from Section 9.4 and use 9.3 as example

Recall: The goal in Machine Learning is to minimize true risk of a predictor $E_{p(x,y)}[l(y, f_\theta(x))]$

- Probabilistic Approach: Use data to estimate $\hat{p}_\theta(x, y)$ that approximates $p(x, y)$, then $f_\theta(x) = \arg\min_{\hat{y} \in \mathcal{Y}} E_{\hat{p}_\theta(y|x)}[l(y, \hat{y})]$

- How to search for optimal $\hat{p}_\theta(y; x)$ or $\hat{p}_\theta(y, x)$?
  - Graphical Models are a language for specifying a family of distributions, D-separation – Conditional Independence define the space of distributions.
  - Searching for $\hat{p}_\theta(x, y)$ requires Estimation Methods
    - MLE is a general rule for estimation of the parameters of a probabilistic model. MLE requires exact data likelihood
      - Many of the times computing the exact data likelihood is intractable then you need the **EM algorithm.** I.e. EM is an approximation algorithm for MLE.

The Probabilistic Model

- ❑ Denote **ALL** observed variables by **X**
- ❑ Denote **ALL** hidden variables by **Z** (also called **H**(hidden), or **L**(latent), or **Y**(output))
  - ○ We only observe the values of **X**
- ❑ According to the model the joint distribution is governed by parameters $p_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta})$
  - ○ NOT conditional dist. $p_\theta(X|Z)$, $p_\theta(Z|X)$
- ❑ Our goal is to implement the MLE rule to learn/estimate $\hat{\theta}^{MLE}$ by maximizing $p(\boldsymbol{X}|\boldsymbol{\theta})$: $\mathrm{p}(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta})$
- ❑ EM necessary when computing $p(\boldsymbol{X}|\boldsymbol{\theta})$ is intractable (Examples)

Unsupervised Naïve Bayes

X = {X_1, …, X_4}

Z = {Y}

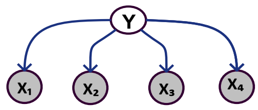
Say we only observe a big corpus of emails but no labels.

We want to estimate

$$\theta_{ij} = p(x_i = 1 \mid Y = j)$$

Tractable Summation

EM not necessary in vanilla model, but with parametric priors



The Probabilistic Model

- ❑ Denote **ALL** observed variables by **X**
- ❑ Denote **ALL** hidden variables by **Z** (also called **H**(hidden), or **L**(latent), or **Y**(output))
  - ○ We only observe the values of **X**
- ❑ According to the model the joint distribution is governed by parameters $p_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta})$
  - ○ NOT conditional dist. $p_\theta(X|Z)$, $p_\theta(Z|X)$
- ❑ Our goal is to implement the MLE rule to learn/estimate $\hat{\theta}^{MLE}$ by maximizing $p(\boldsymbol{X}|\boldsymbol{\theta})$: $\mathrm{p}(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta})$
- ❑ EM Necessary when computing $p(\boldsymbol{X}|\boldsymbol{\theta})$ is intractable (Examples)

Gaussian Mixture Model

$\boldsymbol{Z} = \{\boldsymbol{z_1}, \dots, \boldsymbol{z}_N\}$

$\boldsymbol{X} = \{\boldsymbol{x_1}, \dots \boldsymbol{x}_N\}$

We want to estimate

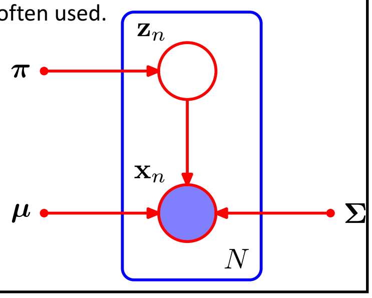
$\boldsymbol{\pi} =$ **Mixture Probabilities**

$\boldsymbol{\mu} = \boldsymbol{Component\ Means}$

$\boldsymbol{\Sigma} =$ **Component Variances**

by maximizing $p(\boldsymbol{X} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$

EM not necessary (in this model), often used.



The EM Algorithm

- Goal is to maximize $p(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_Z p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta})$
- The EM algorithm: Create the following sequence $(\boldsymbol{\theta}^t)$

$$\boxed{\begin{aligned} \boldsymbol{\theta}^0 &\leftarrow \text{Smart or Random Initialization} \\ \boldsymbol{\theta}^{t+1} &\leftarrow \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \underset{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^t)}{\mathrm{E}}[\log(p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta}))] \end{aligned}}$$

- EM update rule guarantees that $p(\boldsymbol{X}|\boldsymbol{\theta}^{t+1}) > \boldsymbol{p}(\boldsymbol{X}|\boldsymbol{\theta}^t)$
  $\Rightarrow$ Convergence to local optima if $p(\boldsymbol{X}|\boldsymbol{\theta})$ is bounded.

Glossary

Z is a random variable.
$p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^t)$ is the posterior distribution over **Z.**
$p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta})$ is a function of **Z** (**X** and $\boldsymbol{\theta}$ are fixed).
$\underset{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^t)}{\mathrm{E}}[\log(p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta}))]$ is a function of $\boldsymbol{\theta}$ alone. *(Also called Q function)*

Proof the following iterations always increase likelihood of data

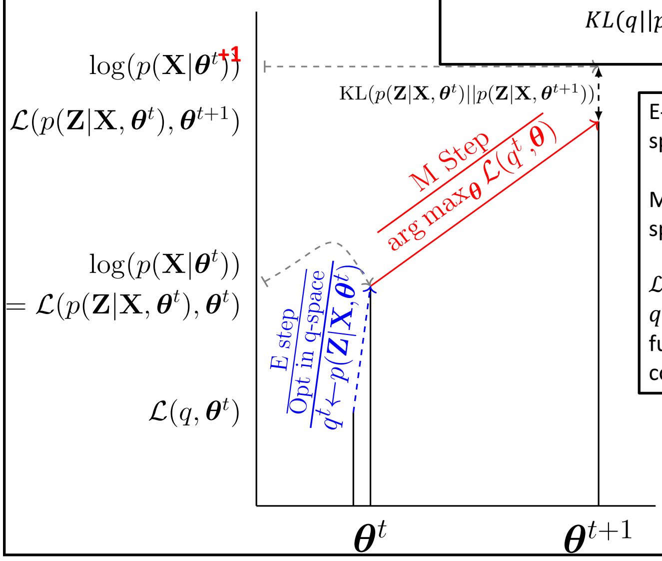
$$\boldsymbol{\theta}^{t+1} \leftarrow \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \underset{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^t)}{\mathrm{E}}[\log(p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta}))]$$

$$KL(q||\,p) \geq 0$$

$$\log(p(\boldsymbol{X} \mid \boldsymbol{\theta})) = \mathcal{L}(q, \boldsymbol{\theta}) + KL(q\,||\,p)$$

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\boldsymbol{z}} \boldsymbol{q}(\boldsymbol{Z}) \log \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{\boldsymbol{q}(\boldsymbol{Z})}$$

$$KL(q||p) = -\sum_{\boldsymbol{z}} q(\boldsymbol{Z}) \log \frac{p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})}{q(\boldsymbol{Z})}$$



E-step is optimization in the space of probabilities

M-step is optimization in space of parameters

$\mathcal{L}(q, \boldsymbol{\theta})$ is a convex function of q. If $p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ is a convex function of $\boldsymbol{\theta}$ then $\mathcal{L}(q, \boldsymbol{\theta})$ is a convex function of $\boldsymbol{\theta}$ as well.

Common Special Case: EM with IID data points

- Assume that two distinct data points (x_i, z_i) and (x_j, z_j) are i.i.d. distributed given $\boldsymbol{\theta}$. This is typically the case when observations are independently generated.

Then $p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta})$

$$= \frac{p(\boldsymbol{X},\boldsymbol{Z} \mid \boldsymbol{\theta})}{\sum_{\boldsymbol{z}} p(\boldsymbol{X},\boldsymbol{Z} \mid \boldsymbol{\theta})} = \frac{\prod_{i=1}^{N} p(x_i,z_i \mid \boldsymbol{\theta})}{\sum_{z} \prod_{i=1}^{N} p(x_i,z_i \mid \boldsymbol{\theta})}$$

$$= \frac{\prod_{i=1}^{N} p(x_i,z_i \mid \boldsymbol{\theta})}{\prod_{i=1}^{N} \sum_{z} p(x_i,z_i \mid \boldsymbol{\theta})} = \prod_{i=1}^{N} p(z_i \mid x_i, \boldsymbol{\theta})$$

And $\underset{p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^t)}{\mathrm{E}}[\log(p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta}))]$ decomposes into

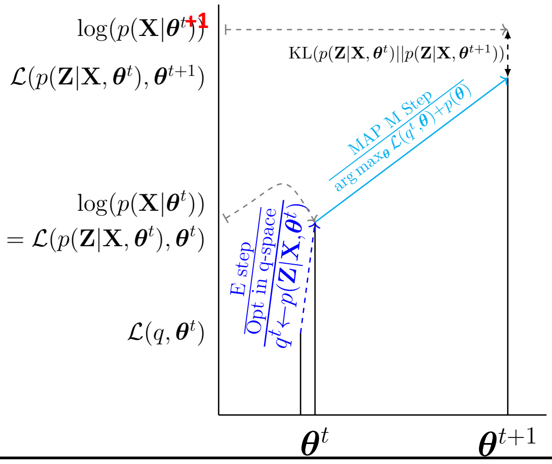
$$\sum_i \underset{p(\boldsymbol{z} \mid \boldsymbol{x_i}, \boldsymbol{\theta}^t)}{\mathrm{E}}[\log(p(\boldsymbol{x_i}, \boldsymbol{z} \mid \boldsymbol{\theta}))]$$

Important Enhancement: EM for MAP Estimation

$\log p(\boldsymbol{\theta} \mid \boldsymbol{X}) = \log p(\boldsymbol{\theta}, \boldsymbol{X}) - \log p(\boldsymbol{X})$

$= \mathcal{L}(q, \boldsymbol{\theta}) + KL(q \,||p) + \log p(\boldsymbol{\theta}) - \log p(X) \geq \mathcal{L}(q, \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \log p(X)$

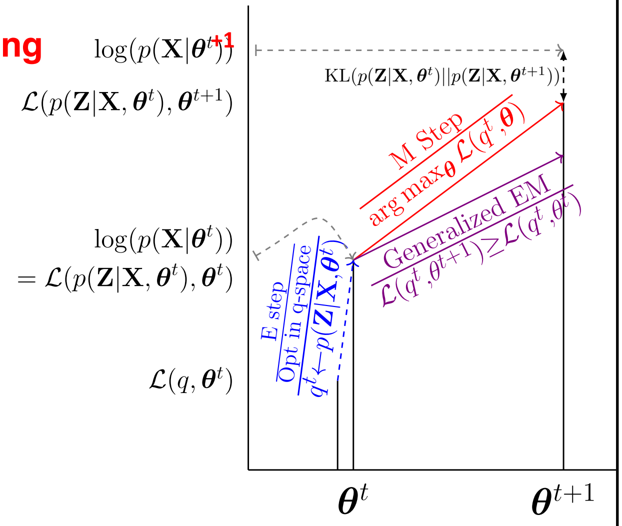
But p(X) is constant so just maximize $\mathcal{L}(q, \boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$ in the M-Step



Potential Generalization: Generalized EM

Instead of **maximizing** the lower bound $\mathcal{L}$ any $\boldsymbol{\theta}$ that **slightly bumps it up** will also do.

For example, Line Search along the gradient will work.



Incremental EM (Motivation)

When $\boldsymbol{p}(\boldsymbol{X},\boldsymbol{Z}\mid\boldsymbol{\theta}) = \prod_{i=1}^{N} p(x_i, z_i \mid \boldsymbol{\theta})$
Then $p(\boldsymbol{Z}\mid\boldsymbol{X},\boldsymbol{\theta}) = \prod_{i=1}^{N} p(z_i|x_i,\boldsymbol{\theta})$ (Slide 8)

$$\Rightarrow \mathcal{L}(q = p(\boldsymbol{Z}\mid\boldsymbol{X},\boldsymbol{\theta}),\boldsymbol{\theta}') = \sum_{\boldsymbol{z}} p(\boldsymbol{Z}\mid\boldsymbol{X},\boldsymbol{\theta}) \log \frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}')}{p(\boldsymbol{Z}\mid\boldsymbol{X},\boldsymbol{\theta})}$$

$$= \sum_i \sum_z p(z|x_i,\boldsymbol{\theta}) \log \frac{p(x_i, z|\boldsymbol{\theta}')}{p(z\,|x_i,\boldsymbol{\theta})}$$

Therefore the objective function decomposes into a sum over N terms.

Furthermore, any value of $\boldsymbol{\theta}$ that globally maximizes $\mathcal{L}(q = p(\boldsymbol{Z}\mid\boldsymbol{X},\boldsymbol{\theta}),\boldsymbol{\theta}')$ subject to $\boldsymbol{\theta}' = \boldsymbol{\theta}$ is a global optima of $p(\boldsymbol{X}|\boldsymbol{\theta})$ **(Why)?**

This suggests possibility for incremental update (Come back to it later)

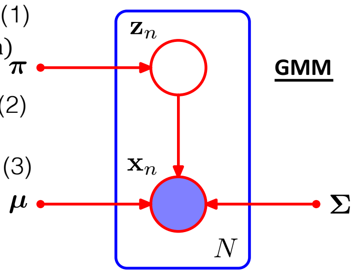
GMM Example: EM, Latent Variables, and Expected Sufficient Statistics

$p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \prod_{\boldsymbol{n=1,k=1}}^{\boldsymbol{N,K}} \pi_k^{z_{nk}} \mathcal{N}(x_n|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})^{z_{nk}}$ (1)

$\Rightarrow E_{q(Z)}[\log p(X, Z \mid \mu, \Sigma, \pi)]$ (q is a general distribution)

$= E_{q(Z)}[\sum_{n\in[N],k\in[K]} z_{nk} (\log \pi_k + \log \mathcal{N}(x_n|\mu_k, \Sigma_k))]$ (2)

$= \sum_{n\in[N],k\in[K]} E_{q(Z)}[z_{nk}] (\log \pi_k + \log \mathcal{N}(x_n|\mu_k, \Sigma_k))$ (3)



Maximize w.r.t. to π by setting:

$$\pi_k \propto \sum_{n\in[N]} E_{q(Z)}[z_{nk}]$$

Maximize wrt μ_k, Σ_k by solving K weighted Least Squares problems.

$$\text{argmax}_{\{\mu_k,\Sigma_k\}} \sum_n - E_{q(Z)}[z_{nk}](\mu_k - x_n)^T \Sigma_k^{-1} (\mu_k - x_n)$$

$$\propto \left(\left[E_{q(Z)}[z_{nk}] \right] x_n, \Sigma_n [E_{q(Z)}[z_{nk}]] |x_n - \mu_n| x_n^T - \mu_n^T \right.$$

The values $E_{q(Z)}[z_{nk}]$ (with q = p(**Z | X**, $\boldsymbol{\theta^t}$)) are called the expected sufficient statistics.
Since in exponential family models the loglikelihood will be linear with respect to these values.

Structured Prediction

(With Special Focus on Sequence Prediction)

* Parts of the following presentation were taken from Professor Mark Dredze's slides

What is Structured Prediction?

- Input: x
  - Typically a structured input
  - Maintain structure of input in x
    - Do not flatten into list of features in an instance
- Output: y
  - y is now from a large set of possible outputs, $\mathcal{Y}$
  - Output space $\mathcal{Y}$ defined based on input
    - Often exponential in size of input

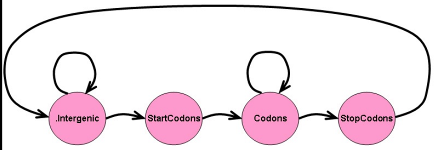
Approaches for structured prediction and why we need special approaches?

- Natural Multi-class algorithms (e.g. Softmax Logistic regression classifiers) don't work with
  - Exponential number of output
  - Outputs defined based on input
- Graphical models for structured prediction
  - Sequences: HMMs and CRFs
- Score based linear models
  - Perceptron, SVM

Examples

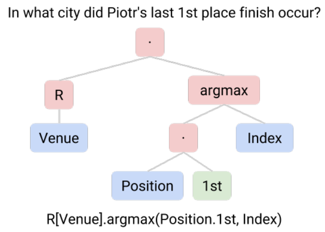
Sequence Prediction

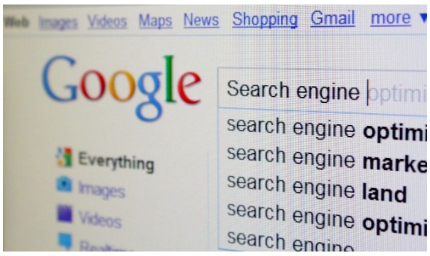
http://cs.wellesley.edu/~cs313/projects/project8/images/OneStrand.jpg



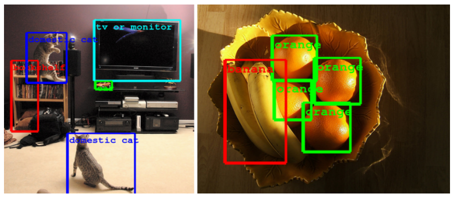
Tree Prediction

http://nlp.stanford.edu/software/sempre/wikitable/images/piotr.png/





Permutation Prediction

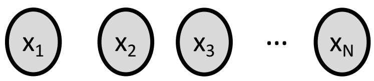


https://3c1703fe8d.site.internapcdn.net/newman/gfx/news/hires/2014/googleteamri.png

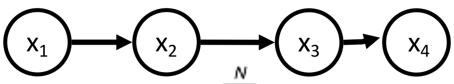
Multi-Object Recognition

Sequential Models

- Simple approach
  - Each event is independent

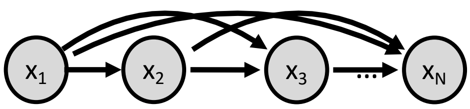


  - $p(x_1, x_2, \dots, x_N) = \prod_{n\in[N]} p(x_n)$
  - Simple, but not very helpful
- **The goldilocks Approach**
  - Markov Assumption



$$p(x_1, x_2 \dots x_N) = \prod_{n=1}^{N} p(x_n \mid x_{n-1})$$

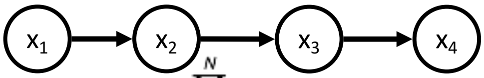
- Complex approach
  - Each event is dependent on previous events



  - $p(x_1, x_2, \dots, x_N) = \prod_{n\in[N]} p(x_n | x_1, \dots, x_{n-1})$
  - Captures dependencies, but way too complex

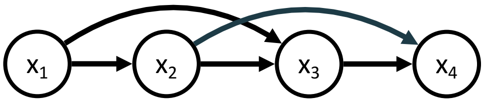
Markov Chains and the Markov Assumption

- First order Markov chain



$$p(x_1, x_2 \ldots x_N) = \prod_{n=1}^{N} p(x_n \mid x_{n-1})$$

- Second order Markov chain

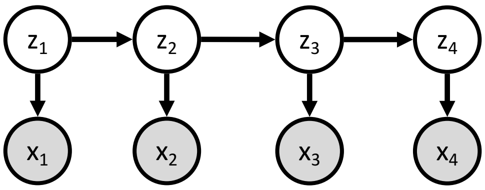


$$p(x_1, x_2 \ldots x_N) = \prod_{n=1}^{N} p(x_n \mid x_{n-1}, x_{n-2})$$

Markov Assumption

- The current state depends on a fixed number of previous states
  - The weather today depends on the past three days, but NOT two weeks ago
- A tractable model that models limited influence of history

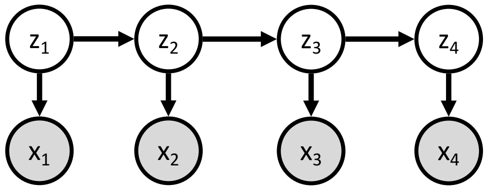
Markov Blankets and Conditional Independence in HMMs



- The Markov blanket for z_n contains z_{n-1}, z_{n+1} and x_n
- The Markov blanket for x_n contains z_n
- Nodes are dependent on a small number of neighbors

Conditional Probability Tables

	$P(z_n=1)$	…
$z_{n-1}=1$		
…		

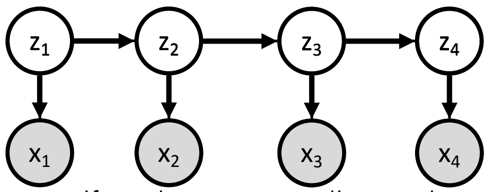


	$P(x_n=1)$	…
$z_n=1$		
…		

Definition: A stationary markov chain is a markov chain where the conditional probability distributions remain the same for each node.

Sequence Models

A HMM is a directed graphical model (BN)



What happens if we have an undirected graphical model?

- Markov Random Field (For modelling $p(\boldsymbol{Z}, \boldsymbol{X})$)
- Conditional Random Fields (For modelling $p(\boldsymbol{Z} \mid \boldsymbol{X})$)
  Go over it in detail later