

Hierarchical Priors

Sargur Srihari

srihari@cedar.buffalo.edu

Topics

1. Hierarchical Priors
2. Bigram model of text dependencies
3. Bag-of-words for Text Classification
4. Latent Dirichlet Allocation

Two extremes and Compromise

1. *Independent BN parameter estimation:*
 - we make strong independence assumptions
 - to decouple estimation of parameters
2. *Shared parameters* is at the other extreme
 - where we force parameters to be identical
- There are situations when neither appropriate
 - Two examples given next
 1. University grades
 2. Word dependency in text domains
- *Compromise* solution is “Soft” parameter sharing is a compromise

Ex 1: University Domain

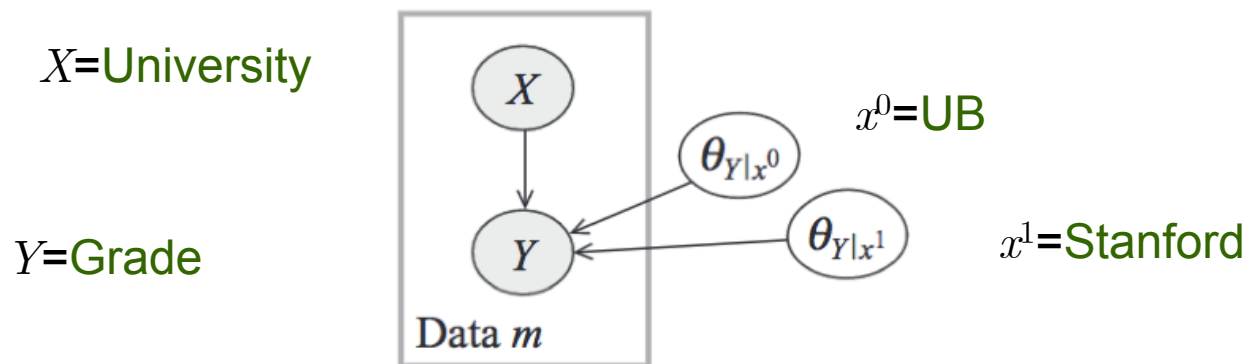
- Suppose we have records of students, classes, teachers from several universities
 - Which may have different properties
 - Engineering CSE, liberal arts CS; different grade scales
- Model over all or each university separately?
 1. Pooling gives more reliable model
 2. Separate allows tailoring to university
 - This doesn't help learn params of other universities
 - Need to learn from scratch that intelligent students tend to get A in easy classes
 - Need to have $P(Y|x)$ similar to each other

Ex 2: Dependencies in text

- Similar problem in learning text dependencies
- Common model is the bigram model
 - Words regarded as forming a Markov chain
 - We have a conditional probability over the next word given the current word: $P(W^{(t+1)} | W^{(t)})$
 - W is a r.v. taking values from dictionary of words
 - Context $W^{(t)}$ can change distribution over $W^{(t+1)}$
 - We still want to share some information across different conditional distributions
 - Probability of “the” should be high in all conditional distributions we learn
 - Need Conditional $P(Y|x)$ similar to each other

Plate model with parameter independence

- This plate model assumes local independence
 - Between two distributions of Y conditioned on values of x



- This plate model is inappropriate for the type of sharing we need

Biasing distributions for similarity

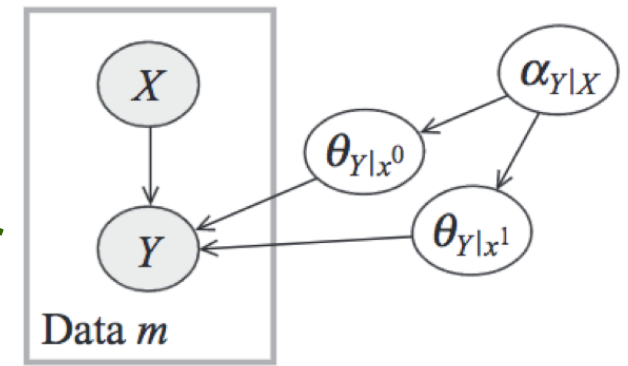
- One way to bias distribution similarity is to have same *prior* over them
 - If prior is very strong it will bias both distributions towards same values
- In the text domain, we want prior to bias both distributions towards giving high probability to frequent words
- How to get such priors
- One solution is to use data to set the prior

Obtaining the prior from data

- Use frequency of words in training set
 - to construct prior where more frequent words have a larger hyperparameter
- Ensures that more frequent words have higher posterior in each of the conditional distributions
 - Even if there are few training samples for that particular conditional distribution
- However it contradicts that a prior is a distribution over parameters before seeing data

A simple hierarchical prior model

- Here we have a variable that is a parent of both $\theta_{Y|x^0}$ and $\theta_{Y|x^1}$
 - Thus the two parameters are no longer independent in the prior and consequently in the posterior
- Intuitively the effect of the prior will be to shift both $\theta_{Y|x^0}$ and $\theta_{Y|x^1}$ to be closer to each other
- Effect of priors diminishes with more data for the different contexts x^0 and x^1
- Thus hierarchical priors are useful for sparse data



Hyperparameter distribution

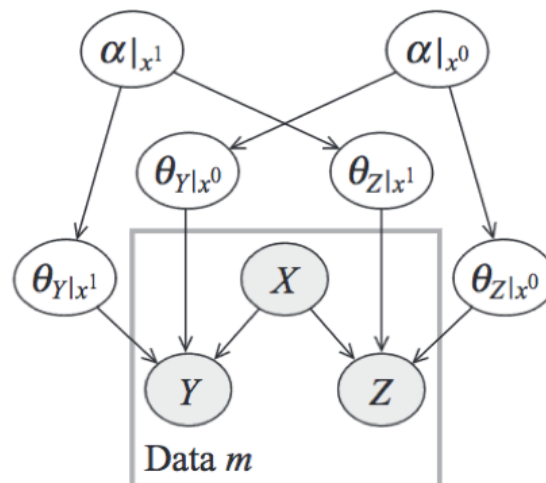
- How to represent distribution over hyperparameters $P(\boldsymbol{\alpha})$
- One option: create a prior where each component α_y is governed by the same distribution, say a Gamma distribution, i.e.,

$$P(\boldsymbol{\alpha}) = \prod_y P(\alpha_y)$$

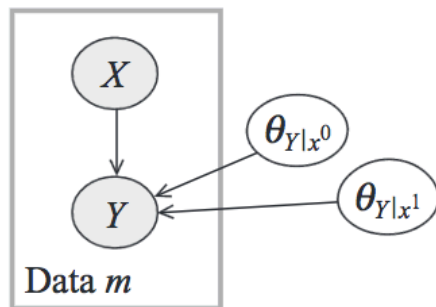
- where $P(\alpha_y) \sim \text{Gamma}(\mu_y)$ is a Gamma distribution with (hyper)hyperparameter μ_y
- Other option: write $\boldsymbol{\alpha}$ as product of equivalent sample size N_0 with a probability distribution p_0
 - First is a Gamma and second is Dirichlet

Dependency between more CPDs

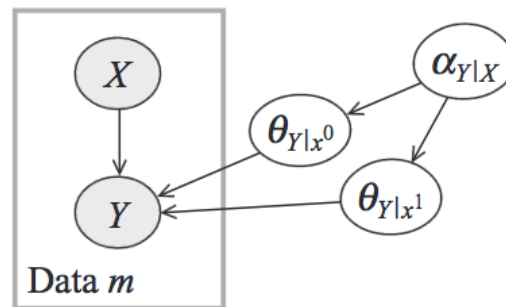
- Can use hierarchical prior with >2 CPDs
- Ex: if we believe that two variables Y and Z depend on X in a similar but not identical way
 - we introduce a common prior on $\theta_{Y|x^0}$ and $\theta_{Z|x^0}$ and similarly another common prior for $\theta_{Y|x^1}$ and $\theta_{Z|x^1}$



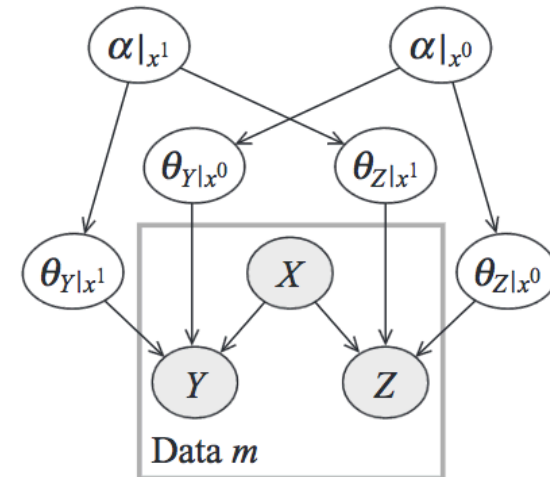
Independent and Hierarchical Priors



A plate model for $p(Y|X)$ under assumption of parameter independence



A plate model for a simple hierarchical prior for the same CPD



A plate model for two CPDs $p(Y|X)$ and $p(Z|X)$ that respond similarly to X

Advantage of Hierarchical Priors

- Provides a flexible language to introduce dependencies in the priors over parameters
- Such dependencies are particularly useful when we have a small no. of samples relevant to each parameter but many such parameters we believe are reasonably similar
- Hierarchical priors spread the effect of observations between parameters with shared hyperparameters