Caltech

Machine Learning & Data Mining

CS/CNS/EE 155

Lecture 7:

Probabilistic Models

Announcements

- Homework 4 released
  - Has coding portion
  - Write clean code!
  - Skeleton code for loading data available on Moodle
    - (You don't have to use it.)

Today

- Basic Probabilistic Models
  - Naïve Bayes
  - Estimation
  - Sampling

- Brief Overview of Advanced Probabilistic Models

- Thursday: Hidden Markov Models in Depth

Generative Probabilistic Models

- Models joint distribution of x and y: $P(x,y)$
- Can make predictions via Bayes Rule:

$$P(y \mid x) = \frac{P(x,y)}{P(x)} = \frac{P(x \mid y)P(y)}{P(x)}$$

Prediction = choose y with maximal P(y|x)

- Can infer marginal distributions:

$$P(y) = \sum_x P(y,x) \qquad P(x) = \sum_y P(y,x)$$

Example

- P(x,y) sums to 1
  - Joint distribution
- P(x=Homework) ??
  - **Answer: 0.5**
  - "Marginalize out the y"

y	x	P(x,y)
Y= SPAM	Help!	0.15
y= NOT	Help!	0.1
y= SPAM	Homework	0.05
y= NOT	Homework	0.45
Y= SPAM	Winner!	0.2
Y= NOT	Winner!	0.05

Margin distribution of P(x)

$$P(x) = \sum_{y} P(y, x)$$

Example #2

- P(x,y) sums to 1
  - Joint distribution

y	x	P(x,y)
Y= SPAM	Help!	0.15
y= NOT	Help!	0.1
y= SPAM	Homework	0.05
y= NOT	Homework	0.45
Y= SPAM	Winner!	0.2
Y= NOT	Winner!	0.05

- P(y=SPAM|x=Help!) ??
  - **Answer: 0.6**
  - P(x,y) = 0.15
  - P(x) = 0.25

$$P(y \mid x) = \frac{P(x,y)}{P(x)} \qquad P(x) = \sum_y P(y,x)$$

Example #3

- P(x,y) sums to 1
  - Joint distribution
- P(x=Help!|y=NOT) ??
  - **Answer: 0.17**
  - P(x,y) = 0.1
  - P(y) = 0.6

y	x	P(x,y)
Y= SPAM	Help!	0.15
y= NOT	Help!	0.1
y= SPAM	Homework	0.05
y= NOT	Homework	0.45
Y= SPAM	Winner!	0.2
Y= NOT	Winner!	0.05

$$P(x \mid y) = \frac{P(x,y)}{P(y)} \qquad P(y) = \sum_x P(y,x)$$

Training

- Goal is to learn P(x,y)
  - What is objective function?

y	x	P(x,y)
Y= SPAM	Help!	0.15
y= NOT	Help!	0.1
y= SPAM	Homework	0.05
y= NOT	Homework	0.45
Y= SPAM	Winner!	0.2
Y= NOT	Winner!	0.05

- **Maximum Likelihood!**

$$\operatorname{argmax} P(S) = \operatorname{argmax} \prod_i P(x_i, y_i)$$

$$= \operatorname{argmin} \sum_i -\log P(x_i, y_i)$$

  - Just frequency counts!
  - 6 parameters

$$S = \{(x_i, y_i)\}_{i=1}^{N}$$

Training

- Goal is to learn P(x,y)
  - What is objective function?

- **Maximum Likelihood!**

$$\operatorname{argmax} P(S) = \operatorname{argmax} \prod_i P(x_i, y_i)$$

$$= \operatorname{argmin} \sum_i -\log P(x_i, y_i)$$

y	x	P(x,y)
Y= SPAM	Help!	0.15
y= NOT	Help!	0.1
y= SPAM	Homework	0.05
y= NOT	Homework	0.45
Y= SPAM	Winner!	0.2
Y= NOT	Winner!	0.05

Interpretation: Given model structure, find that parameterization that best explains data

$\}_{i=1}^{N}$

Training Derivation

- Define: $$P(x,y) = \frac{w_{x,y}}{\sum_{x',y'} w_{x',y'}}$$

Just a re-parameterization

$$\operatorname{argmin} \sum_i -\log P(x_i, y_i) = \underset{w}{\operatorname{argmin}} \sum_i \left[-\log w_{x_i,y_i} + \log \sum_{x',y'} w_{x',y'} \right]$$

\# training examples (x,y)

$$\partial_{w_{x,y}} = -\frac{N_{x,y}}{w_{x,y}} + \frac{N}{\sum_{x',y'} w_{x',y'}} \rightarrow \frac{N_{x,y}}{N} = \frac{w_{x,y}}{\sum_{x',y'} w_{x',y'}} \rightarrow P(x,y) = \frac{N_{x,y}}{N}$$

Frequency of (x,y) in training set!

Regularization

- Hallucinate data!

Prior Probability of observing (x,y)

$$P(x,y) = \frac{N_{x,y} + \lambda P_{x,y}}{N + \lambda}$$

Regularization Strength

y	x	P(x,y)
Y= SPAM	Help!	0.15
y= NOT	Help!	0.1
y= SPAM	Homework	0.05
y= NOT	Homework	0.45
Y= SPAM	Winner!	0.2
Y= NOT	Winner!	0.05

- aka: “pseudo counts”

Generative vs Discriminative

- Generative models
  - Models both y AND x
  - $P(x,y)$

What are Benefits and Drawbacks?

- Discriminative models
  - Models y GIVEN x
  - $P(y|x)$
  - E.g., Logistic Regression

Generative vs Discriminative

- Generative:
  - 6 parameters in example
  - Can sample P(x,y)
  - Prediction via Bayes Rule
    - Tolerates missing data
- Discriminative:
  - 3 parameters in example
  - Can only sample P(y|x)
  - Directly models prediction task
    - Cannot naturally tolerate missing data

y	x	P(x,y)
Y= SPAM	Help!	0.15
y= NOT	Help!	0.1
y= SPAM	Homework	0.05
y= NOT	Homework	0.45
Y= SPAM	Winner!	0.2
Y= NOT	Winner!	0.05

Discriminative Models Make Better Predictions

- Directly learn to optimize prediction goal:
  - Aka: directly learn: $P(y \mid x)$
  - E.g., minimize log-loss

- Generative Models require combining multiple estimated values:

$$P(y \mid x) = \frac{P(x, y)}{P(x)}$$

What if there are so many different x that P(x) underflows?

  - Training objective does not maximize accuracy.

Generative Models are Joint Models

- Fully specify probability distribution of P(x,y)
- Can draw samples from P(x,y)
  - R = uniform([0,1])
  - If(R < 0.15)
    - x=help!, y=SPAM
  - Elseif(R < 0.25)
    - x=help!, y=NOT
  - …

Built-in function in python, Matlab, etc.

y	x	P(x,y)
Y= SPAM	Help!	0.15
y= NOT	Help!	0.1
y= SPAM	Homework	0.05
y= NOT	Homework	0.45
Y= SPAM	Winner!	0.2
Y= NOT	Winner!	0.05

Generative Models can Tolerate Missing Values

- We can model the probability of missing feature value
  - We will see this specifically for Naïve Bayes.

- Discriminative models cannot tolerate missing values
  - If you don't observe an input feature, you lose all guarantees

Generative Models are more Elegant?

- Many people find generative models more elegant
- Tell a "complete" story about the data
- Useful if we can't decide what is the prediction task a priori
- E.g., train model first, pick what is the y later

Naïve Bayes

Modeling a Feature Vector

- Single y
  - (e.g., binary)
- Vector of x (D-dimensional)
  - Simplest case, each x^d binary
  - E.g., presence/absence of word

- Model P(x,y)

Example

- Binary y
- 2 binary x's

- "Probability table"

- **What's wrong with this approach?**

y	x^1=Winner!	x^2=Homework	P(x,y)
SPAM	1	1	0.01
NOT	1	1	0.01
SPAM	0	1	0.03
NOT	0	1	0.35
SPAM	1	0	0.25
NOT	1	0	0.05
SPAM	0	0	0.2
NOT	0	0	0.1

Example

- Binary y
- 2 binary x's

- "Probability table"

y	x^1=Winner!	x^2=Homework	P(x,y)
SPAM	1	1	0.01
NOT	1	1	0.01
SPAM	0	1	0.03
NOT	0	1	0.35
SPAM	1	0	0.25
NOT	1	0	0.05
			0.2
			0.1

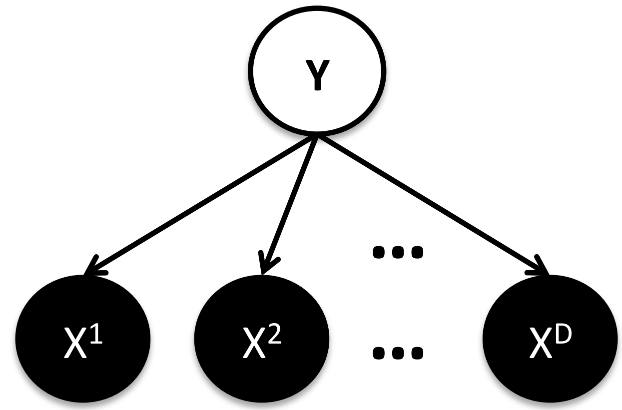
- **Wha**

this

Model Complexity is Exponential w.r.t. the length of x!

Naïve Bayes Formulation

- Posits a generating model:
  - Single y
  - Multiple x features
  - **Only keep track of:**
    - **P(y), P(x^d|y)**



Graphical Model Diagram

$$P(x,y) = P(x \mid y)P(y) = P(y)\prod_d P(x^d \mid y)$$

Each x^d is conditionally independent given y.
"Naïve" independence assumption!

Why is Naïve Bayes Convenient?

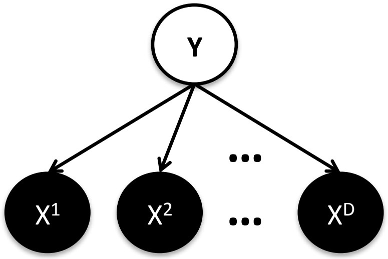
- Compact representation
- Easy to compute any quantity
  - $P(y|x)$, $P(x^d|y)$, …
- Easy to estimate model components
  - $P(y)$, $P(x^d|y)$
- Easy to sample
- Easy to deal with missing values

Example Model (Discrete)

- Each x^d binary
  - E.g., presence or absence of word

P(x|y)

	x^1=Homework	x^2=Winner!
y=SPAM	$P(x^1\|y)=0.2$	$P(x^2\|y)=0.5$
y=NOT	$P(x^1\|y)=0.6$	$P(x^2\|y)=0.1$



P(y)

	P(y)
y=SPAM	0.7
y=NOT	0.3

$$P(x,y) = P(x \mid y)P(y) = P(y)\prod_d P(x^d \mid y)$$

Example Model (Discrete)

- Each x^d binary
  - E.g., presence or absence of word

P(x|y)

	x^1=Homework	x^2=Winner!
y=SPAM	$P(x^1\|y)=0.2$	$P(x^2\|y)=0.5$
y=NOT	$P(x^1\|y)=0.6$	$P(x^2\|y)=0.1$

Model Complexity is Linear w.r.t. the length of x!

P(y)

	P(y)
y=SPAM	0.7
y=NOT	0.3

$$P(x,y) = P(x \mid y)P(y) = P(y)\prod_d P(x^d \mid y)$$

Making Predictions

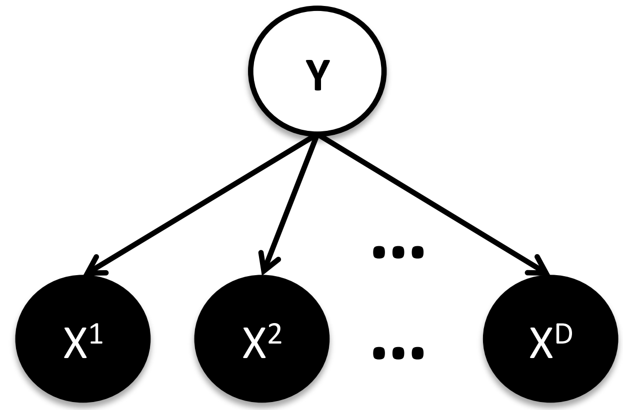
$$P(y \mid x) = \frac{P(x,y)}{P(x)}$$

$$= \frac{P(x \mid y)P(y)}{P(x)}$$

$$= \frac{P(y)}{P(x)} \prod_d P(x^d \mid y)$$

$$\propto P(y) \prod_d P(x^d \mid y)$$

Model components we keep track of.



Graphical Model Diagram

Example Prediction

- Suppose:

$$P(y=1)=0.3 \qquad P(x \mid y=1)=0.05$$

$$P(y=-1)=0.7 \qquad P(x \mid y=-1)=0.001$$

- Then:

$$P(y=1 \mid x)=\frac{0.3*0.05}{0.3*0.05+0.7*0.001}\approx 0.96$$

$$P(y \mid x) \propto P(y)\prod_d P(x^d \mid y) = P(y)P(x \mid y)$$

Example Prediction #2

- What if we want to compute: $P(x^1 \mid x^{2:D}, y)$

- **Simple!** $P(x^1 \mid y)$

- It's an explicitly defined model component:

$$P(x, y) = P(x \mid y)P(y) = P(y)\prod_d P(x^d \mid y)$$

Example Prediction #3

- What if we want to compute: $P(x^1 \mid x^{2:D})$

$$P(x^1 \mid x^{2:D}) = \frac{P(x)}{P(x^{2:D})} = \frac{\sum_y P(y)P(x \mid y)}{\sum_y P(y)P(x^{2:D} \mid y)}$$

"Marginalizing out the y"

Why is the numerator smaller than the denominator?

$$P(x,y) = P(x \mid y)P(y) = P(y)\prod_d P(x^d \mid y)$$

Marginalization in Matrix Form

Often faster than writing for loops!

O

	x^1=Homework	x^2=Winner!
y=SPAM	$P(x^1=1 \mid y)=0.2$	$P(x^2=1 \mid y)=0.5$
y=NOT	$P(x^1=1 \mid y)=0.6$	$P(x^2=1 \mid y)=0.1$

P

	P(y)
y=SPAM	0.7
y=NOT	0.3

- Compute $P(x^d=1)$:

$$P(x^d = 1) = \left[O^T P\right]_d$$

d-th row

$$P(x^d = 1) = \sum_y P(x^d = 1 \mid y) P(y)$$

Missing Values

- What if we don't observe x^2?
- Predict $P(y\text{=SPAM}|x^1)$

We can marginalize out the missing values!

$$P(y \mid x^1) = \sum_{x^{2:D}} P(y, x^{2:D} \mid x^1) = \sum_{x^{2:D}} \frac{P(x,y)}{P(x^1)}$$

How to efficiently sum over multiple missing values?

	x^1=Homework	x^2=Winner!
y=SPAM	$P(x^1=1\|y)=0.2$	$P(x^2=1\|y)=0.5$
y=NOT	$P(x^1=1\|y)=0.6$	$P(x^2=1\|y)=0.1$

	P(y)
y=SPAM	0.7
y=NOT	0.3

Conditional Independence to the Rescue!

$$P(y \mid x^1) = \sum_{x^{2:D}} P(y, x^{2:D} \mid x^1) = \sum_{x^{2:D}} \frac{P(x,y)}{P(x^1)}$$

From previous slide

$$P(x,y) = P(y) \prod_d P(x^d \mid y)$$

Definition of Naïve Bayes

$$\sum_{x^{2:D}} P(x,y) = P(y) \sum_{x^{2:D}} \prod_d P(x^d \mid y)$$

$$= P(y) P(x^1 \mid y) \prod_{d \in [2,D]} \sum_{x^d} P(x^d \mid y)$$

Swap Product & Sum due to independence!

$$= P(y) P(x^1 \mid y)$$

Marginalizes to 1!

Intuition

- Consider the case of 3 variables in x:

$$\sum_{x^{2:D}} P(x,y) = P(y)\sum_{x^{2:D}}\prod_{d} P(x^d \mid y) = P(y)P(x^1 \mid y)\prod_{d\in[2,D]}\sum_{x^d} P(x^d \mid y) = P(y)P(x^1 \mid y)$$

$$= \sum_{x^2\in\{0,1\}}\sum_{x^3\in\{0,1\}} P(x^2 \mid y)P(x^3 \mid y)$$

$$\begin{aligned}
&= P(x^2=0 \mid y)P(x^3=0 \mid y) + P(x^2=0 \mid y)P(x^3=1 \mid y)\\
&\quad + P(x^2=1 \mid y)P(x^3=0 \mid y) + P(x^2=1 \mid y)P(x^3=1 \mid y)
\end{aligned}$$

$$= \big(P(x^2=0 \mid y) + P(x^2=1 \mid y)\big)\big(P(x^3=0 \mid y) + P(x^3=1 \mid y)\big)$$

$$= 1$$

Intuition

- Consider the case of 3 variables in x:

$$\sum_{x^{2:D}} P(x,y) = P(y) \sum_{x^{2:D}} \prod_{d} P(x^d \mid y) = P(y) P(x^1 \mid y) \prod_{d\in[2,D]} \sum_{x^d} P(x^d \mid y)$$

$$= \sum_{x^2\in\{0,1\}} \sum_{x^3\in\{0,1\}} P(x^2 \mid y) P(x^3 \mid y) = \sum_{x^2\in\{0,1\}} \sum_{x^3\in\{0,1\}} P(x^2 \mid y) P(x^3 \mid y)$$

$$= P(x^2=0 \mid y) P(x^3=0 \mid y) + P(x^2=0 \mid y) P(x^3=1 \mid y)$$
$$+ P(x^2=1 \mid y) P(x^3=0 \mid y) + P(x^2=1 \mid y) P(x^3=1 \mid y)$$

$$= \Big(P(x^2=0 \mid y) + P(x^2=1 \mid y)\Big)\Big(P(x^3=0 \mid y) + P(x^3=1 \mid y)\Big)$$

One Empirical Comparison

MODEL	1ST	2ND	3RD	4TH	5TH	6TH	7TH	8TH	9TH	10TH
BST-DT	0.580	0.228	0.160	0.023	0.009	0.000	0.000	0.000	0.000	0.000
RF	0.390	0.525	0.084	0.001	0.000	0.000	0.000	0.000	0.000	0.000
BAG-DT	0.030	0.232	0.571	0.150	0.017	0.000	0.000	0.000	0.000	0.000
SVM	0.000	0.008	0.148	0.574	0.240	0.029	0.001	0.000	0.000	0.000
ANN	0.000	0.007	0.035	0.230	0.606	0.122	0.000	0.000	0.000	0.000
KNN	0.000	0.000	0.000	0.009	0.114	0.592	0.245	0.038	0.002	0.000
BST-STMP	0.000	0.000	0.002	0.013	0.014	0.257	0.710	0.004	0.000	0.000
DT	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.616	0.291	0.089
LOGREG	0.000	0.000	0.000	0.000	0.000	0.000	0.040	0.312	0.423	0.225
NB	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.030	0.284	0.686

- Measure how frequently each model places
  - 1st, 2nd, 3rd, etc.
- Only generative model (Naïve Bayes) is in last place

“An Empirical Comparison of Supervised Learning Algorithms”
Caruana, Niculescu-Mizil, ICML 2006

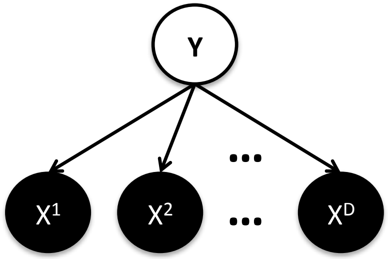
Training

- Maximum Likelihood of Training Set:

$$\underset{}{\operatorname{argmax}} P(S) = \operatorname{argmax} \prod_i P(x_i, y_i) \qquad S = \{(x_i, y_i)\}_{i=1}^N$$

$$= \operatorname{argmin} \sum_i -\log P(x_i, y_i)$$

  – Subject to Naïve Bayes assumption on structure of P(x,y)



Only need to estimate P(y) and each P(x^d|y)!

$$P(x,y) = P(x \mid y)P(y) = P(y)\prod_d P(x^d \mid y)$$

Just Counting!

$$P(y = SPAM) = \frac{N_{y=SPAM}}{N}$$

Frequency of SPAM documents in training set

$$P(x^1 = 1 \mid y = SPAM) = \frac{N_{y=SPAM \wedge x^1=1}}{N_{y=SPAM}}$$

Frequency of word x1 appearing in SPAM documents in training set

Regularization

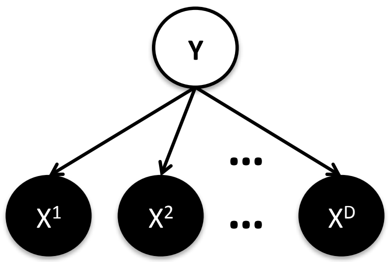
- Add "pseudo counts"
  - aka hallucinate some data

$$P(y = SPAM) = \frac{N_{y=SPAM} + \lambda P_{y=SPAM}}{N + \lambda}$$

Often just set pseudo counts to uniform distribution!

$$P(x^1 = 1 \mid y = SPAM) = \frac{N_{y=SPAM \wedge x^1=1} + \lambda P_{y=SPAM \wedge x^1=1}}{N_{y=SPAM} + \lambda}$$

Sampling

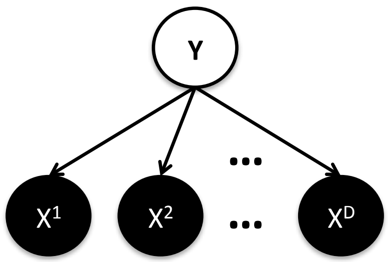


- Can sample from distribution
  - Definition of Generative Model
- Can draw samples from P(x,y)
  - First sample y:
    - Random uniform variable R
    - Set y=SPAM if R < P(y=SPAM) & y=NOT otherwise
  - Then sample each x^d:
    - Sample uniform variable R
    - Set x^d=1 if R < P(x^d=1|y) & x^d=0 otherwise

Built-in function in python, Matlab, etc.

Sampling Example

- Sample P(y)
  - R = 0.5, so set y = SPAM
- Sample $P(x^1|y=SPAM)$
  - R = 0.1, so set $x^1 = 1$
- Sample $P(x^1|y=SPAM)$
  - R = 0.9, so set $x^2 = 0$

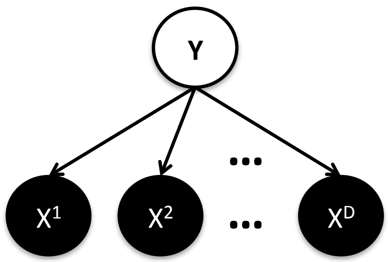


Can be done in either order

	x^1=Homework	x^2=Winner!
y=SPAM	$P(x^1=1\|y)=0.2$	$P(x^2=1\|y)=0.5$
y=NOT	$P(x^1=1\|y)=0.6$	$P(x^2=1\|y)=0.1$

	P(y)
y=SPAM	0.7
y=NOT	0.3

Sampling Example #2



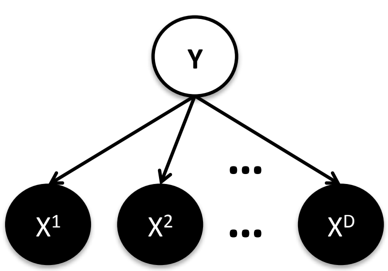
- Sample P(y)
  - R = 0.9, so set y = NOT
- Sample $P(x^1 | y{=}NOT)$
  - R = 0.5, so set $x^1 = 1$
- Sample $P(x^1 | y{=}NOT)$
  - R = 0.05, so set $x^2 = 1$

	x^1=Homework	x^2=Winner!
y=SPAM	$P(x^1=1 \mid y)=0.2$	$P(x^2=1 \mid y)=0.5$
y=NOT	$P(x^1=1 \mid y)=0.6$	$P(x^2=1 \mid y)=0.1$

	P(y)
y=SPAM	0.7
y=NOT	0.3

Recap: Naïve Bayes

- Probabilistic Generative Model
- Make strong independence assumptions
  - Compact representation
  - Easy to train
  - Easy to compute various probabilities
  - Not the most accurate for standard prediction



$$P(x,y) = P(x \mid y)P(y) = P(y)\prod_d P(x^d \mid y)$$

Invent Your Own Model

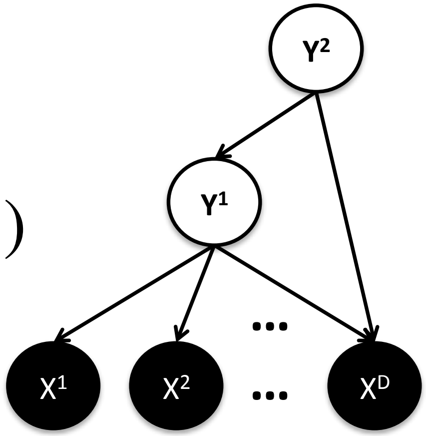
- Naïve Bayes is a special case of Bayesian Network

- Here's another one I just made up:

$$P(x,y) = P(x \mid y)P(y)$$

$$= P(x \mid y)P(y^1 \mid y^2)P(y^2)$$

$$= P(y^1 \mid y^2)P(y^2)P(x^D \mid y^1,y^2)\prod_{d\in[1,D-1]} P(x^d \mid y^1)$$

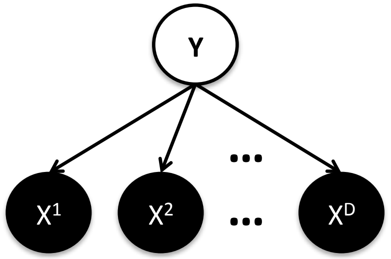


Some Other Probabilistic Models

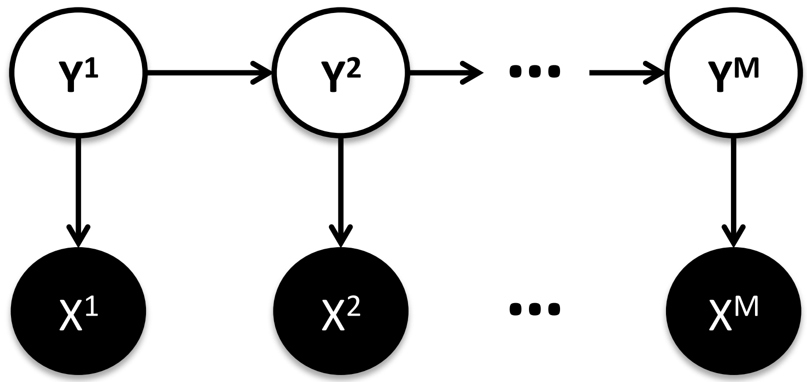
Gaussian Naïve Bayes

- Same independence structure as Naïve Bayes
  - But probability functions are now Gaussians
    - (Instead of discrete lookup tables.)
  - y is binary: $P(y)$ the same
  - Each x^d is continuous:

$$P(x^d \mid y) \sim N(\mu_{d,y}, \sigma)$$



Hidden Markov Models



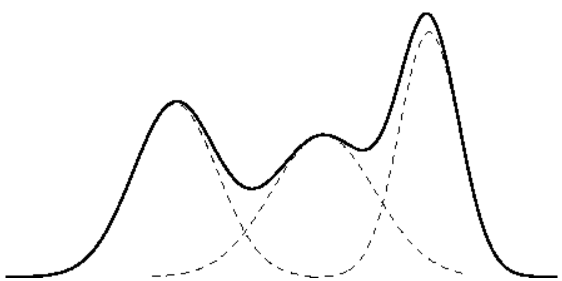
- Generative model of sequences

$$P(x,y) = P(y^1)P(x^1 \mid y^1)\prod_{j=2}^{M} P(y^j \mid y^{j-1})P(x^j \mid y^j)$$

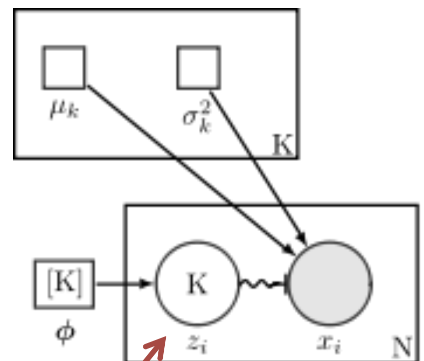
- (focus of next lecture)

(Gaussian) Mixture Models

- Each data point is associated with a membership to a Gaussian distribution
  – Denoted by z variable
- 1D Example with 3 Gaussians



K Gaussian Distributions



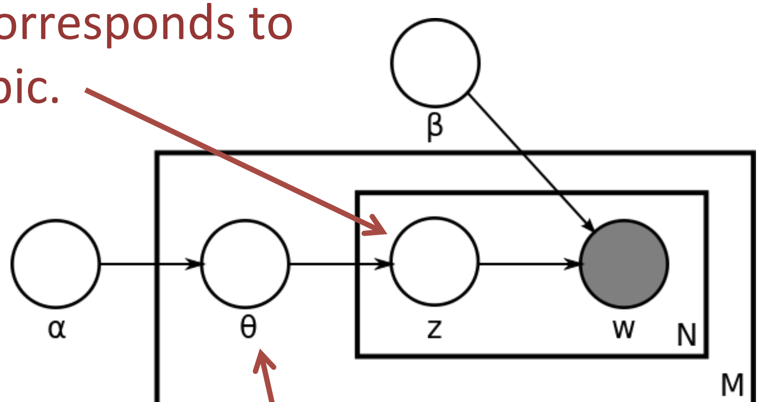
N Data Points

Membership variable per data point

Topic Models
(Latent Dirichlet Allocation)

- Posits that documents can represented as a mixture of topics.
  - K topics, choose K a priori
- Posits that topics can be represented as a mixture of words

Each word corresponds to a specific topic.



Training set: M documents, each with N words.

Topic mixture of document.

Example: LDA analysis of Sarah Palin's emails

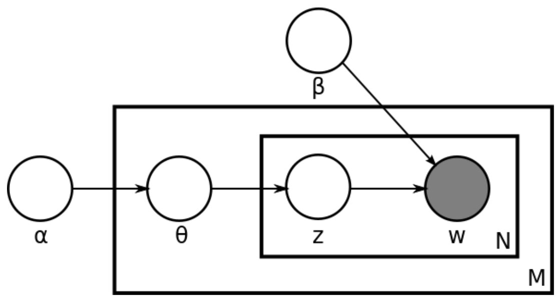
(Disclaimer: this was the top result of Google Search "LDA example")

- **Topics**:
  - **Trig/Family/Inspiration**: family, web, mail, god, son, from, congratulations, children, life, child, down, trig, baby, birth, love, you, syndrome, very, special, bless, old, husband, years, thank, best, …
  - **Wildlife/BP Corrosion**: game, fish, moose, wildlife, hunting, bears, polar, bear, subsistence, management, area, board, hunt, wolves, control, department, year, use, wolf, habitat, hunters, caribou, program, denby, fishing, …
  - **Energy/Fuel/Oil/Mining:** energy, fuel, costs, oil, alaskans, prices, cost, nome, now, high, being, home, public, power, mine, crisis, price, resource, need, community, fairbanks, rebate, use, mining, villages, …
  - **Gas**: gas, oil, pipeline, agia, project, natural, north, producers, companies, tax, company, energy, development, slope, production, resources, line, gasline, transcanada, said, billion, plan, administration, million, industry, …
  - **Education/Waste**: school, waste, education, students, schools, million, read, email, market, policy, student, year, high, news, states, program, first, report, business, management, bulletin, information, reports, 2008, quarter, …
  - **Presidential Campaign/Elections**: mail, web, from, thank, you, box, mccain, sarah, very, good, great, john, hope, president, sincerely, wasilla, work, keep, make, add, family, republican, support, doing, p.o, …

http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/

Example: LDA analysis of Sarah Palin's emails
(Disclaimer: this was the top result of Google Search "LDA example")

- Presidential Campaign
- Wildlife




```
We understand that you have been discussed as a possible choice for
the Vice Presidency.

As people who support the democratic process and care about
protecting our wildlife for future generations, we want you to know
that we don't believe people in our states would
vote for you for any office if they knew your record on these issues.

It is troubling that you are now working to deny more than 50,000
Alaskans a vote on aerial killing of wolves and bears with
legislation now being considered in the Alaska legislature.
```

Deep Belief Networks

http://gitxiv.com/posts/jG46ukGod8R7Rdtud/a-neural-algorithm-of-artistic-style

Recap: Generative Probabilistic Models

- Quantifies Uncertainty
  - Can tolerate missing values

- Model represents a “summary” of the data
  - Fit model parameters to data
  - Can use for inspection

- Not trained to optimize prediction accuracy

Next Lecture

- Hidden Markov Models in depth
  - Sequence Modeling
  - Requires Dynamic Programming
  - Implement aspects of HMMs in homework

- Recitation Thursday:
  - Recap of Dynamic Programming (for HMMs)