6.S064 Spring 2013
Introduction to Machine Learning

(slides without jokes...)

6.S064 Spring 2013
Introduction to Machine Learning

http://web.mit.edu/subjectevaluation/

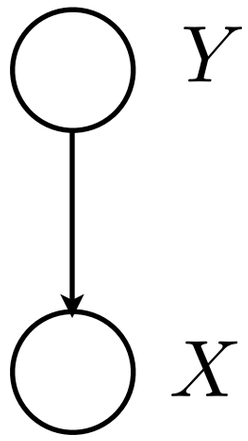
What do you need to know?

- Discriminative models and methods
  - linear classifiers, perceptron, max-margin hyperplane
  - non-linear classifiers, feature mappings, kernels
  - linear/non-linear regression
- Concepts
  - regularization, model selection, generalization
- Generative models and methods
  - mixture models, the EM-algorithm
  - hidden markov models
  - bayesian networks
- Decisions and actions
  - reinforcement learning

What do you need to know?

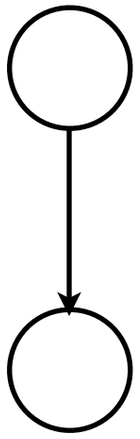
- Discriminative models and methods
  - linear classifiers, perceptron, max-margin hyperplane
  - non-linear classifiers, feature mappings, kernels
  - linear/non-linear regression
- Concepts
  - regularization, model selection, generalization
- Generative models and methods
  - mixture models, the EM-algorithm
  - hidden markov models
  - bayesian networks
- Decisions and actions
  - reinforcement learning

Mixture models example



Mixture models example

$P(Y = y)$



$Y \in \{0, 1\}$

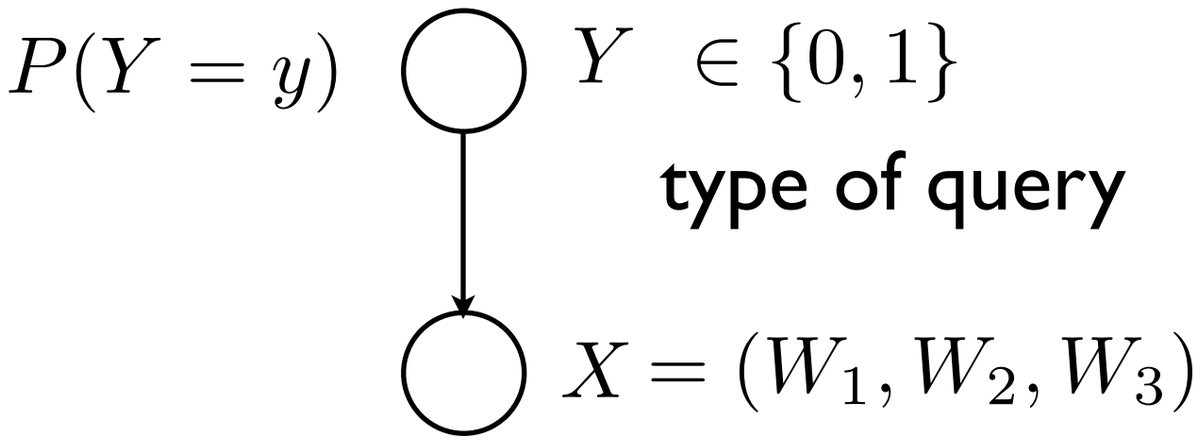
type of query

$X = (W_1, W_2, W_3)$

keywords

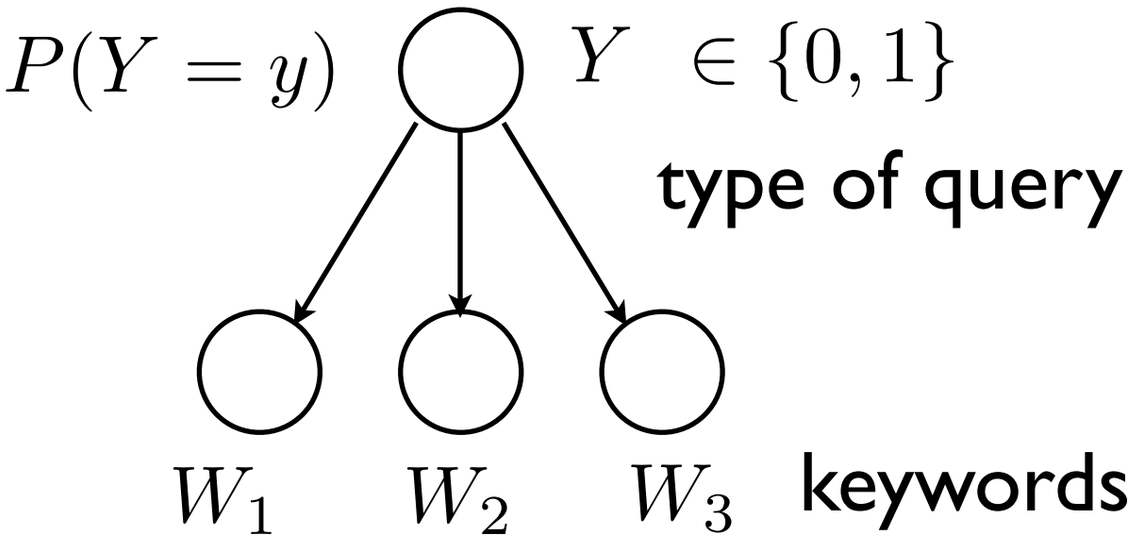
$P(X = x|Y = y)$

Mixture models example



y	(w_1, w_2, w_3)
1	What is SVM?
0	Mother's day flowers
1	6.S064 final exams
	...

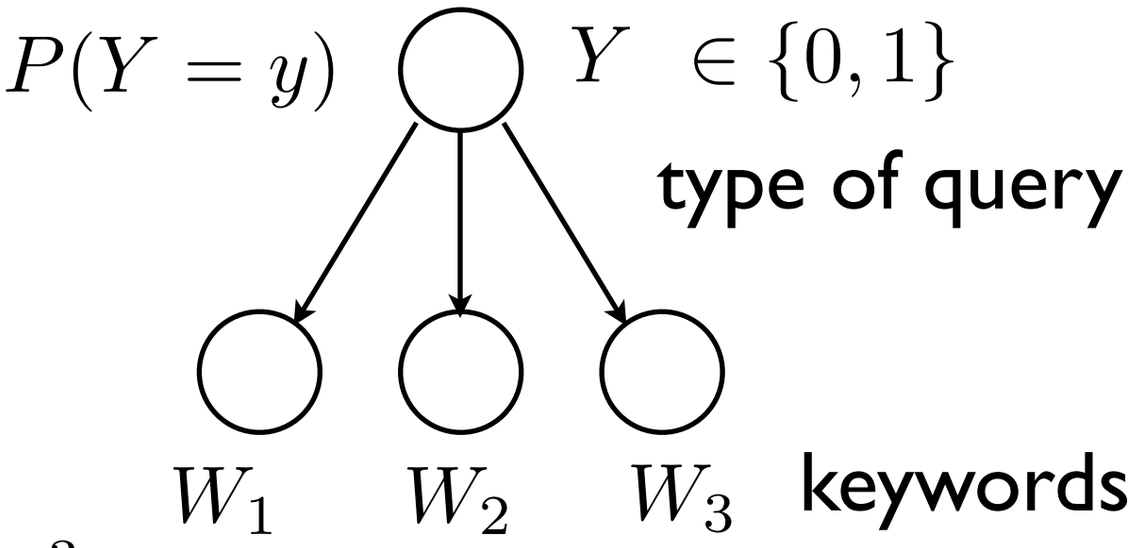
Mixture models example



(assumption!!)

y	(w_1, w_2, w_3)
1	What is SVM?
0	Mother's day flowers
1	6.S064 final exams
	...

Mixture models example



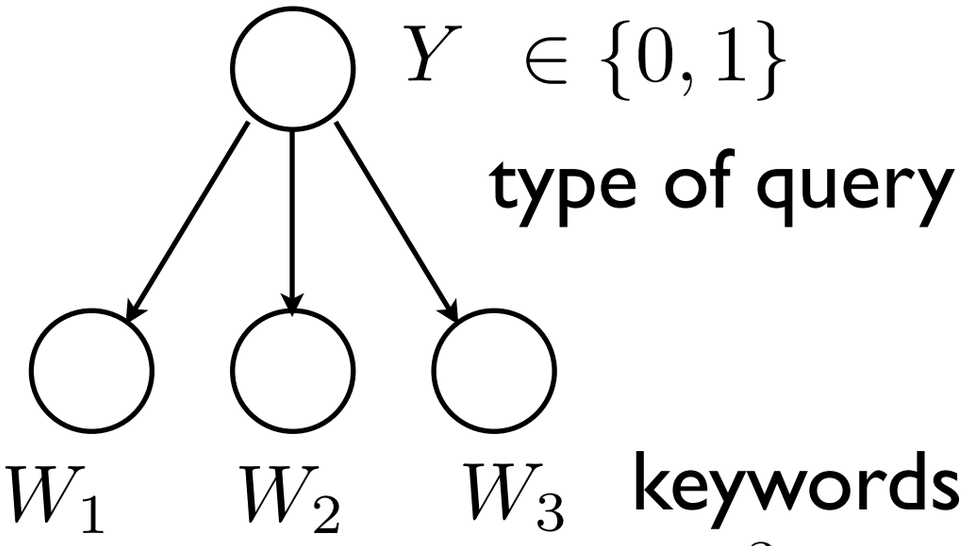
(assumption!!)

$$\prod_{i=1}^{3} P(W_i = w_i | Y = y)$$

y	(w_1, w_2, w_3)
1	What is SVM?
0	Mother's day flowers
1	6.S064 final exams
	...

Mixture models example

$$P(Y = y) = \theta(y)$$

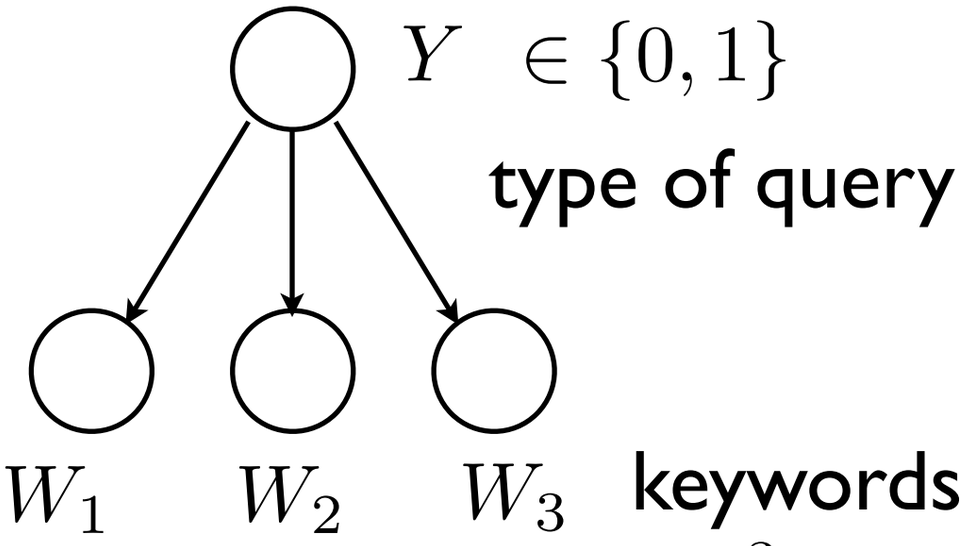


$$\prod_{i=1}^{3} P(W_i = w_i | Y = y) = \prod_{i=1}^{3} \theta_i(w_i | y)$$

y	(w_1, w_2, w_3)
1	What is SVM?
0	Mother's day flowers
1	6.S064 final exams
	...

Mixture models example

$$P(Y = y) = \theta(y)$$



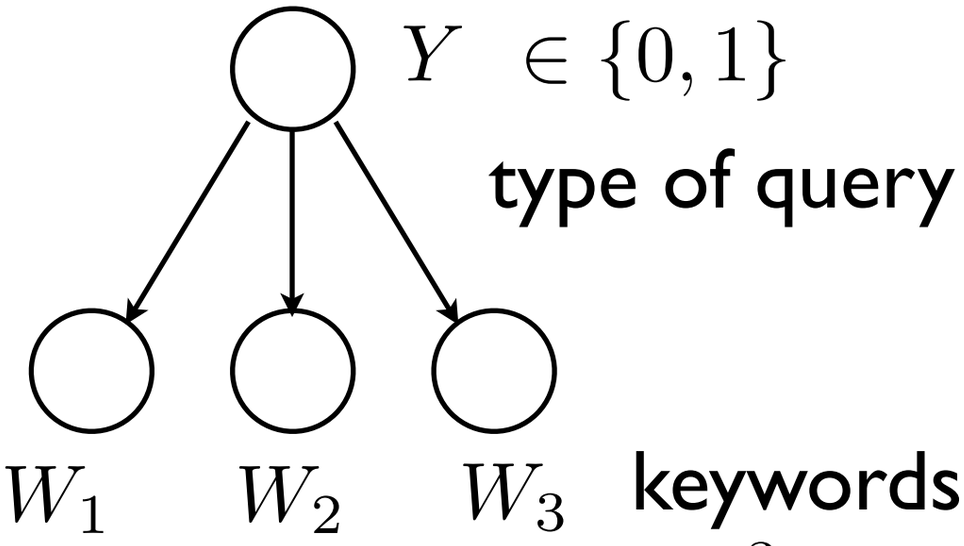
$$\prod_{i=1}^{3} P(W_i = w_i | Y = y) = \prod_{i=1}^{3} \theta_i(w_i|y)$$

$$\hat{\theta}_1(What|1) =$$

y	(w_1, w_2, w_3)
1	What is SVM?
0	Mother's day flowers
1	6.S064 final exams
	...

Mixture models example

$$P(Y = y) = \theta(y)$$



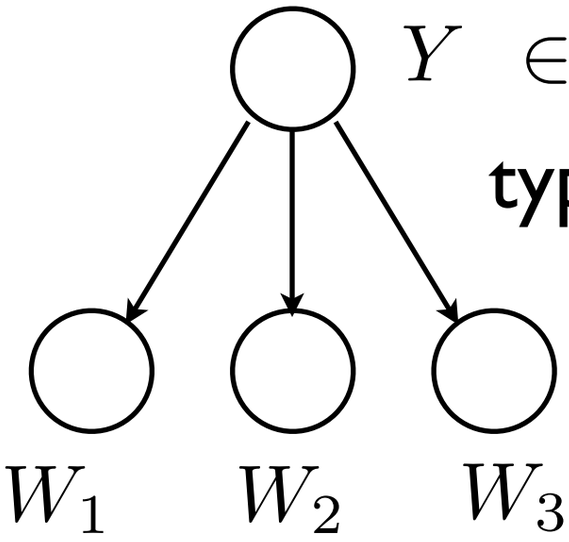
$$\prod_{i=1}^{3} P(W_i = w_i | Y = y) = \prod_{i=1}^{3} \theta_i(w_i|y)$$

$$\hat{\theta}_1(What|1) = \frac{1}{2}$$

y	(w_1, w_2, w_3)
1	What is SVM?
0	Mother's day flowers
1	6.S064 final exams
	...

Mixture models example

$$P(Y = y) = \theta(y)$$



$$\prod_{i=1}^{3} P(W_i = w_i | Y = y) = \prod_{i=1}^{3} \theta_i(w_i | y)$$

$$\hat{\theta}_1(What|1) = \frac{1}{2}$$

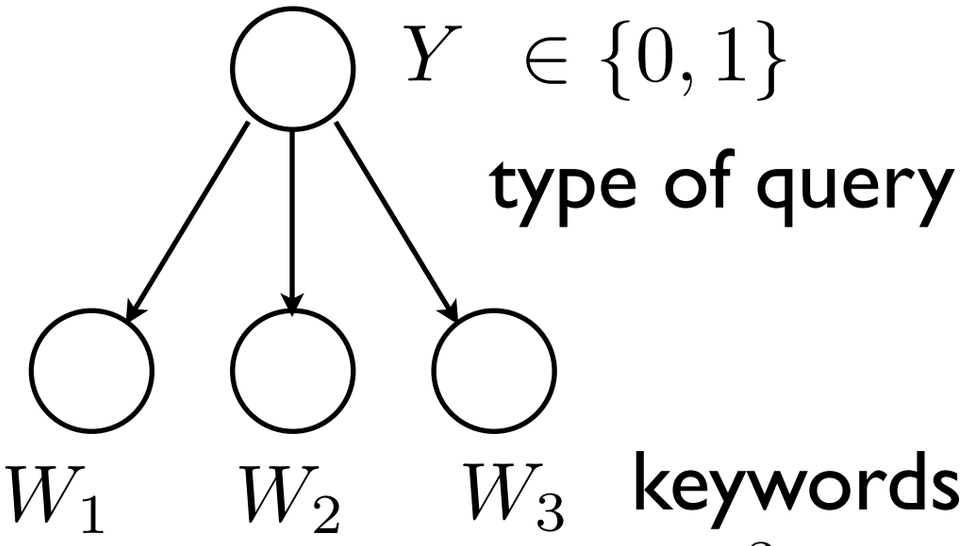
$$\hat{\theta}_1(What|0) = 0$$

...

y	(w_1, w_2, w_3)
1	What is SVM?
0	Mother's day flowers
1	6.S064 final exams
	...

Mixture models example

$$P(Y = y) = \theta(y)$$

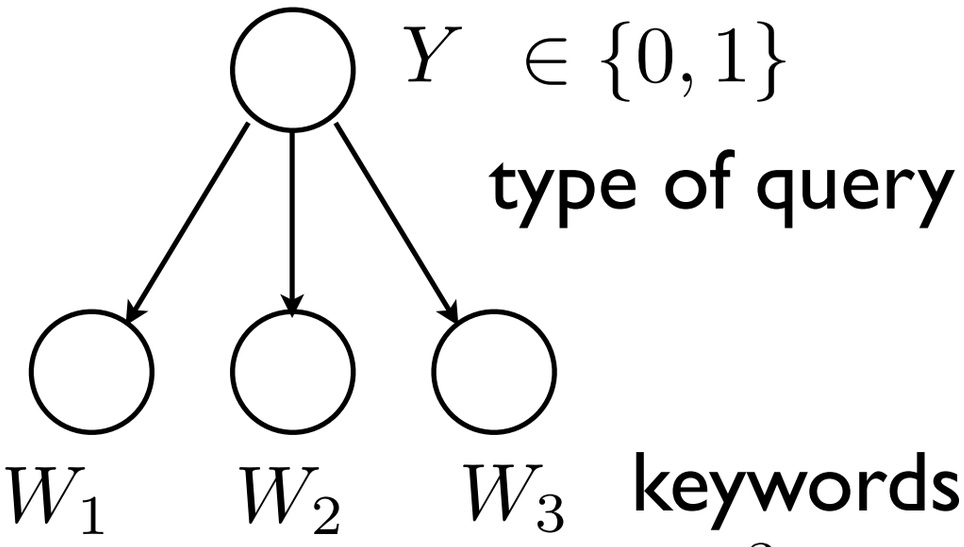


$$\prod_{i=1}^{3} P(W_i = w_i | Y = y) = \prod_{i=1}^{3} \theta_i(w_i | y)$$

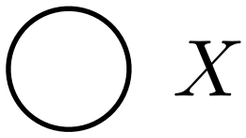
y	(w_1, w_2, w_3)
?	What is SVM?
?	Mother's day flowers
?	6.S064 final exams
	...

Mixture models example

$$P(Y = y) = \theta(y)$$



$\Rightarrow$



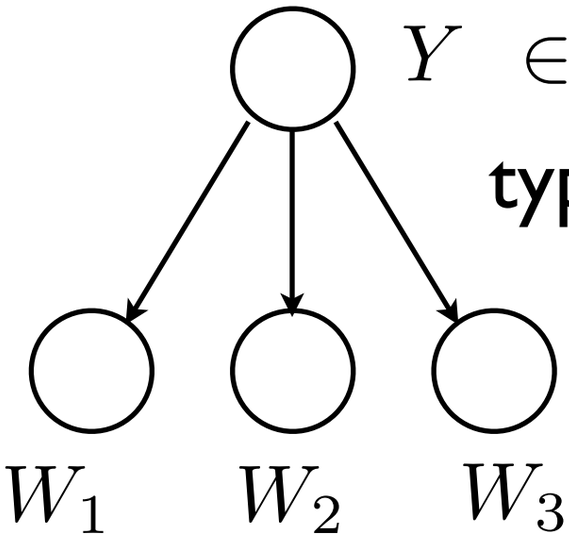
$$P(X = (w_1, w_2, w_3)|\theta)$$

$$\prod_{i=1}^{3} P(W_i = w_i|Y = y) = \prod_{i=1}^{3} \theta_i(w_i|y)$$

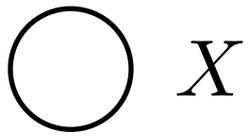
y	(w_1, w_2, w_3)
?	What is SVM?
?	Mother's day flowers
?	6.S064 final exams
	...

Mixture models example

$P(Y = y) = \theta(y)$



$\Rightarrow$



$$P(X = (w_1, w_2, w_3)|\theta)$$

$$\prod_{i=1}^{3} P(W_i = w_i|Y = y) = \prod_{i=1}^{3} \theta_i(w_i|y)$$

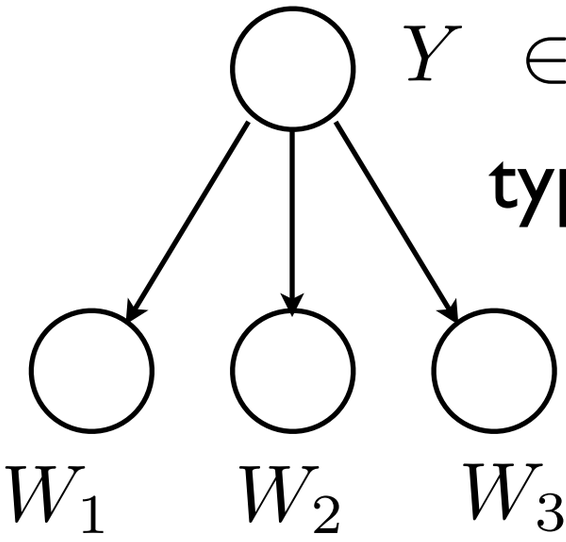
$$= \sum_y \theta(y) \prod_{i=1}^{3} \theta_i(w_i|y)$$

mixture!

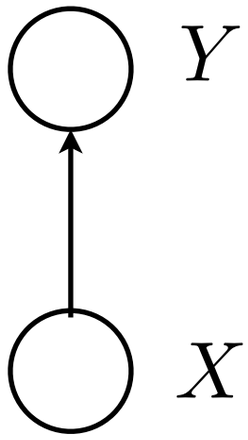
y	(w_1, w_2, w_3)
?	What is SVM?
?	Mother's day flowers
?	6.S064 final exams
	...

Mixture models example

$$P(Y = y) = \theta(y)$$



$$\Rightarrow$$



$$\prod_{i=1}^{3} P(W_i = w_i | Y = y) = \prod_{i=1}^{3} \theta_i(w_i|y)$$

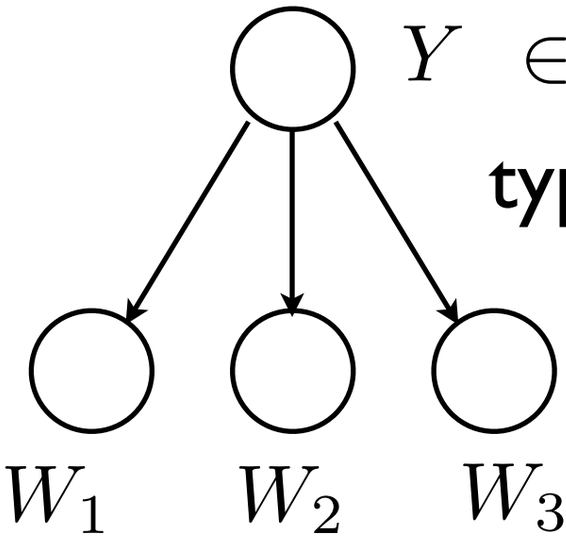
$$P(X = (w_1, w_2, w_3)|\theta) = \sum_y \theta(y) \prod_{i=1}^{3} \theta_i(w_i|y)$$

mixture!

y	(w_1, w_2, w_3)
?	What is SVM?
?	Mother's day flowers
?	6.S064 final exams
	...

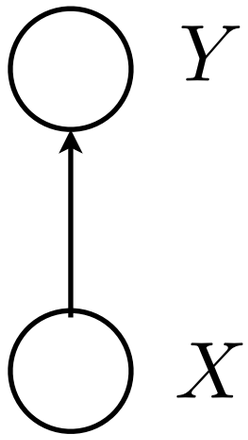
Mixture models example

$$P(Y = y) = \theta(y)$$



$$\Rightarrow$$

$$P(Y = y | X = (w_1, w_2, w_3), \theta)$$



$$P(X = (w_1, w_2, w_3) | \theta)$$

$$\prod_{i=1}^{3} P(W_i = w_i | Y = y) = \prod_{i=1}^{3} \theta_i(w_i | y)$$

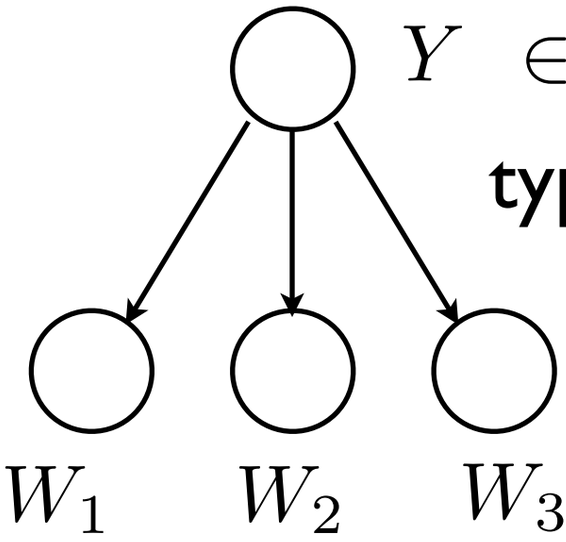
$$= \sum_y \theta(y) \prod_{i=1}^{3} \theta_i(w_i | y)$$

mixture!

y	(w_1, w_2, w_3)
?	What is SVM?
?	Mother's day flowers
?	6.S064 final exams
	...

Mixture models example

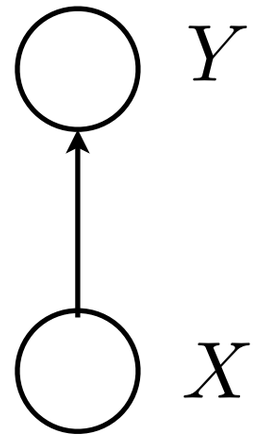
$$P(Y = y) = \theta(y)$$



$$\prod_{i=1}^{3} P(W_i = w_i | Y = y) = \prod_{i=1}^{3} \theta_i(w_i|y)$$

$$P(Y = y | X = (w_1, w_2, w_3), \theta)$$

=



$$P(X = (w_1, w_2, w_3)|\theta)$$

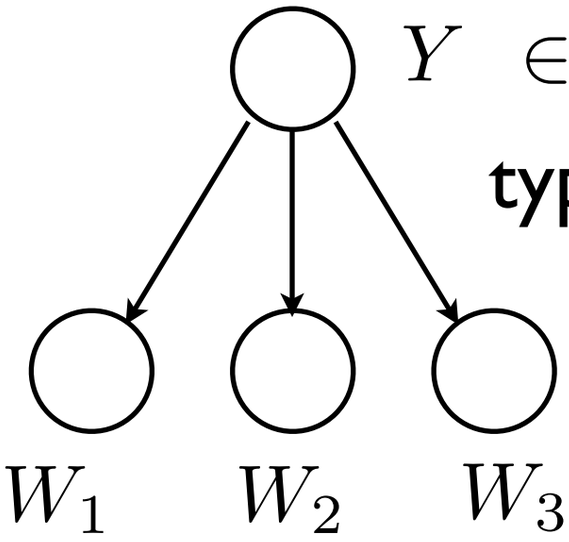
$$= \sum_{y} \theta(y) \prod_{i=1}^{3} \theta_i(w_i|y)$$

mixture!

y	(w_1, w_2, w_3)
?	What is SVM?
?	Mother's day flowers
?	6.S064 final exams
	...

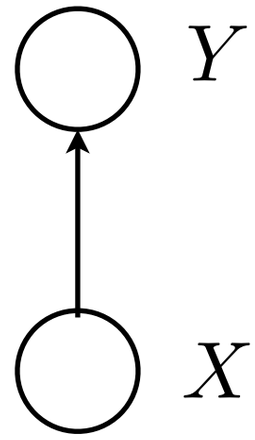
Mixture models example

$$P(Y = y) = \theta(y)$$



$$\prod_{i=1}^{3} P(W_i = w_i | Y = y) = \prod_{i=1}^{3} \theta_i(w_i|y)$$

$$P(Y = y | X = (w_1, w_2, w_3))$$



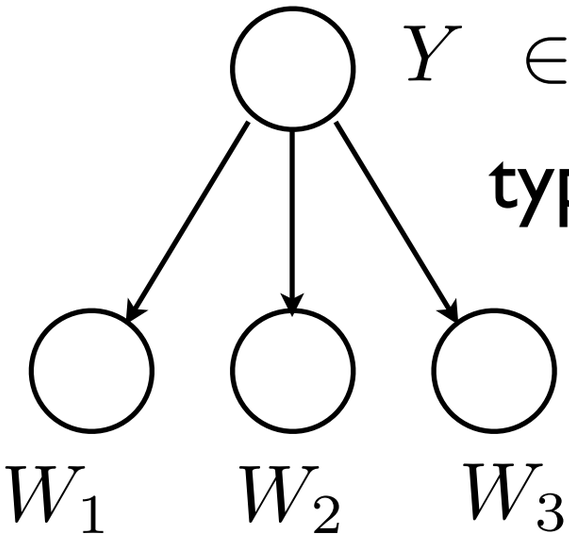
$$\hat{P}(X = (w_1, w_2, w_3))$$

empirical distribution

y	(w_1, w_2, w_3)
?	What is SVM?
?	Mother's day flowers
?	6.S064 final exams
	...

Mixture models: the EM algorithm

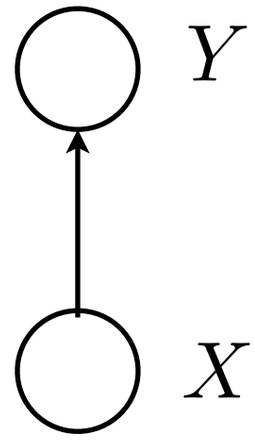
$$P(Y = y) = \theta(y)$$



$$\prod_{i=1}^{3} P(W_i = w_i | Y = y) = \prod_{i=1}^{3} \theta_i(w_i|y)$$

$$P(Y = y | X = (w_1, w_2, w_3))$$

$\approx$



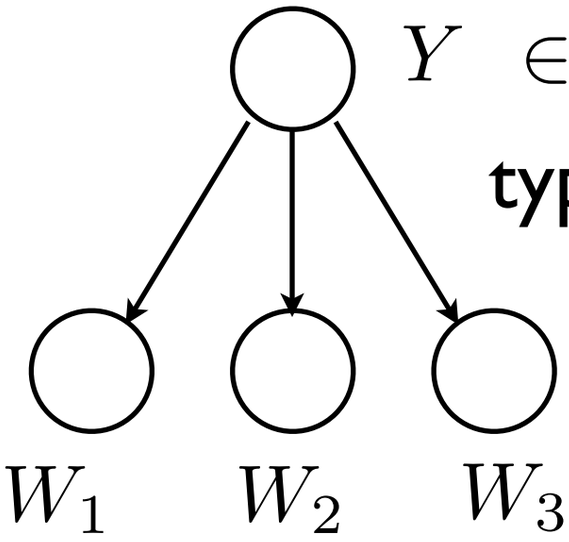
$$\hat{P}(X = (w_1, w_2, w_3))$$

empirical distribution

y	(w_1, w_2, w_3)
?	What is SVM?
?	Mother's day flowers
?	6.S064 final exams
	...

Mixture models: the EM algorithm

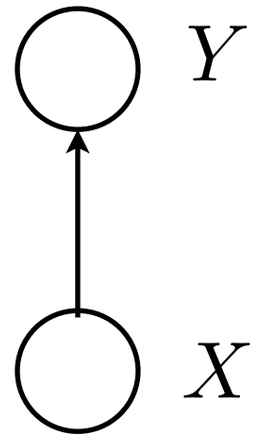
$$P(Y = y) = \theta(y)$$



$$\prod_{i=1}^{3} P(W_i = w_i | Y = y) = \prod_{i=1}^{3} \theta_i(w_i|y)$$

$$\approx$$

$$P(Y = y | X = (w_1, w_2, w_3), \theta)$$



$$\hat{P}(X = (w_1, w_2, w_3))$$

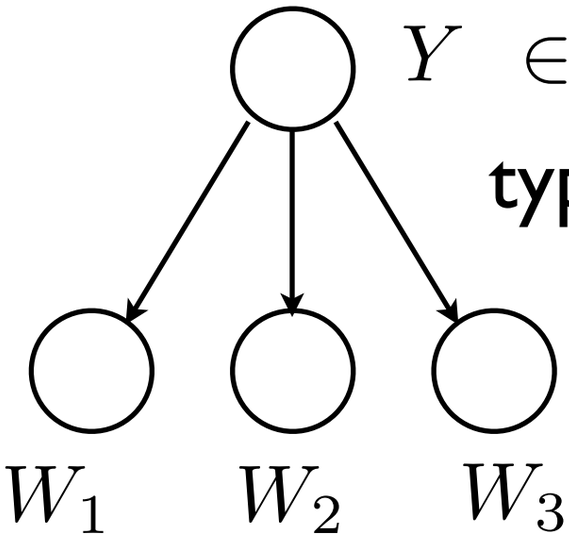
empirical distribution

y	(w_1, w_2, w_3)
?	What is SVM?
?	Mother's day flowers
?	6.S064 final exams
	...

E-step

Mixture models: the EM algorithm

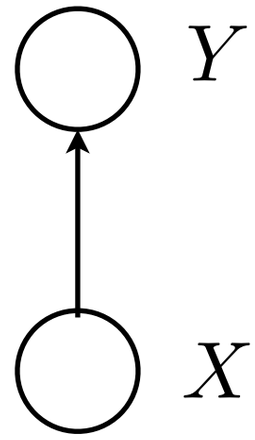
$$P(Y = y) = \theta(y)$$



$$\prod_{i=1}^{3} P(W_i = w_i | Y = y) = \prod_{i=1}^{3} \theta_i(w_i|y)$$

$$\approx$$

$$P(Y = y | X = (w_1, w_2, w_3), \theta)$$



$$\hat{P}(X = (w_1, w_2, w_3))$$

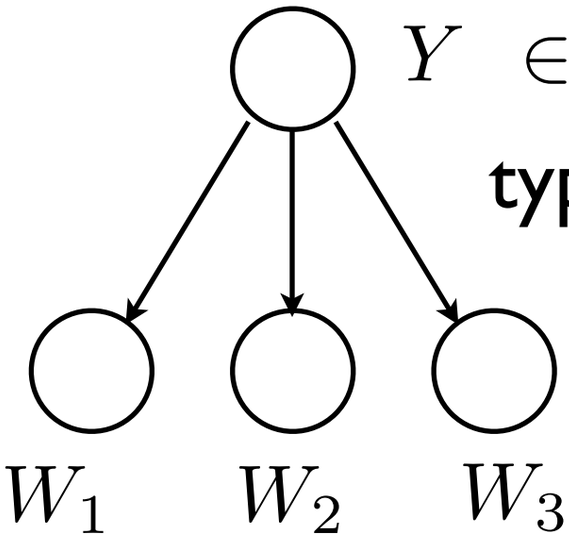
empirical distribution

$P(Y = y \mid X = (w_1, w_2, w_3), \theta)$	y	(w_1, w_2, w_3)
0.7	1	What is SVM?
0.3	0	What is SVM?
0.1	1	Mother's day flowers
0.9	0	Mother's day flowers
		...

E-step

Mixture models: the EM algorithm

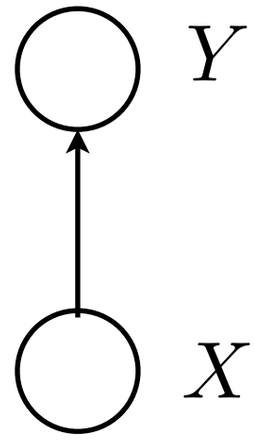
$$P(Y = y) = \theta(y)$$



$$\prod_{i=1}^{3} P(W_i = w_i | Y = y) = \prod_{i=1}^{3} \theta_i(w_i|y)$$

$$\approx$$

$$P(Y = y | X = (w_1, w_2, w_3), \theta)$$



$$\hat{P}(X = (w_1, w_2, w_3))$$

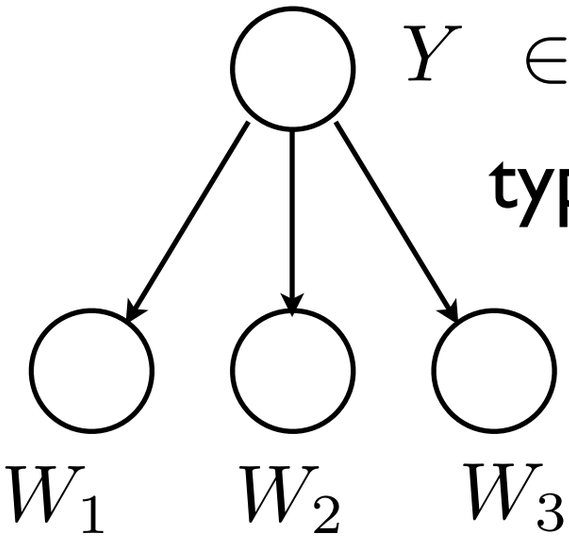
empirical distribution

$P(Y = y \mid X = (w_1, w_2, w_3), \theta)$	y	(w_1, w_2, w_3)
0.7	1	What is SVM?
0.3	0	What is SVM?
0.1	1	Mother's day flowers
0.9	0	Mother's day flowers
		...

E-step

Mixture models: the EM algorithm

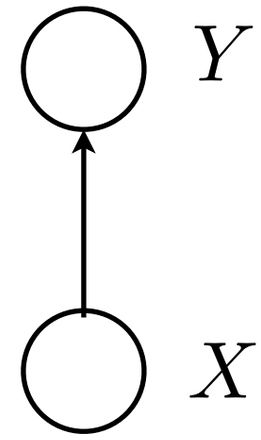
$$P(Y = y) = \theta(y)$$



$$\prod_{i=1}^{3} P(W_i = w_i | Y = y) = \prod_{i=1}^{3} \theta_i(w_i|y)$$

$\approx$

$$P(Y = y | X = (w_1, w_2, w_3), \theta)$$



$$\hat{P}(X = (w_1, w_2, w_3))$$

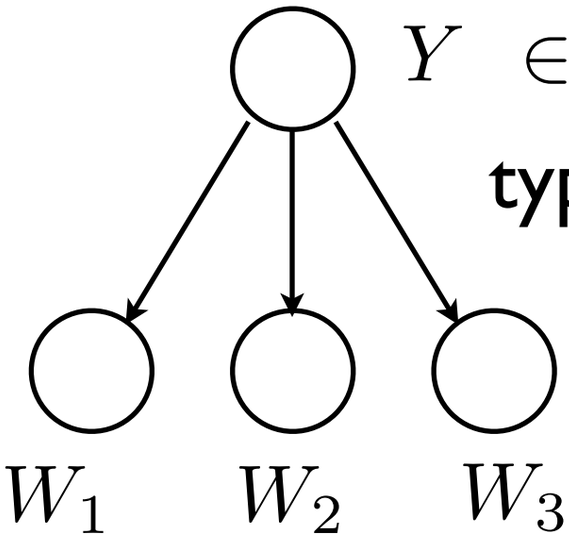
empirical distribution

	y	(w_1, w_2, w_3)
0.7	1	What is SVM?
0.3	0	What is SVM?
0.1	1	Mother's day flowers
0.9	0	Mother's day flowers
		...

M-step

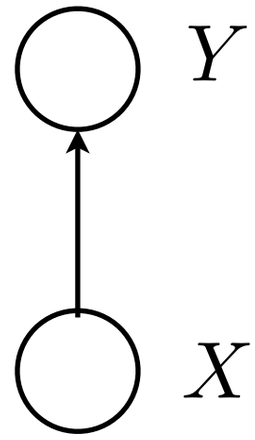
Mixture models: the EM algorithm

$$P(Y = y) = \theta(y)$$



$$\approx$$

$$P(Y = y | X = (w_1, w_2, w_3), \theta)$$



$$\hat{P}(X = (w_1, w_2, w_3))$$

empirical distribution

$$\prod_{i=1}^{3} P(W_i = w_i | Y = y) = \prod_{i=1}^{3} \theta_i(w_i | y)$$

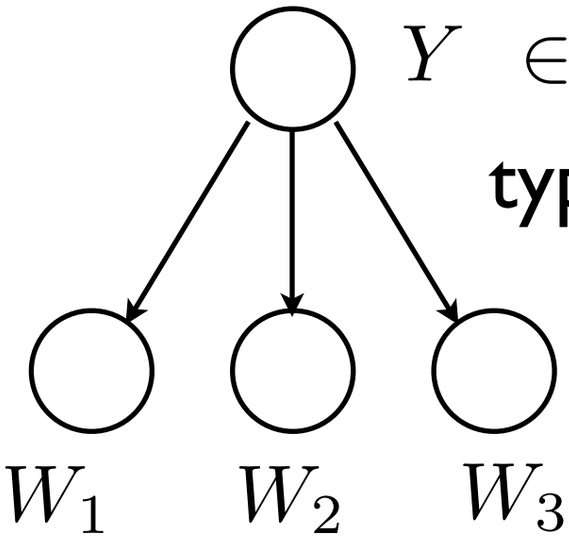
$$\hat{\theta}_1(What|1) =$$

y	(w_1, w_2, w_3)	
0.7	1	What is SVM?
0.3	0	What is SVM?
0.1	1	Mother's day flowers
0.9	0	Mother's day flowers
		...

M-step

Mixture models: the EM algorithm

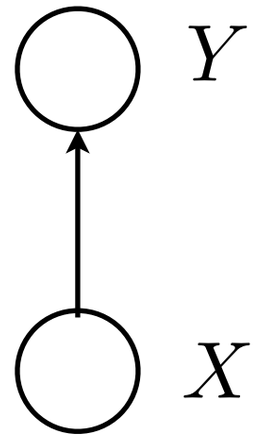
$$P(Y = y) = \theta(y)$$



$$\prod_{i=1}^{3} P(W_i = w_i | Y = y) = \prod_{i=1}^{3} \theta_i(w_i|y)$$

$$\hat{\theta}_1(What|1) = \frac{0.7}{0.7 + 0.1}$$

$$P(Y = y | X = (w_1, w_2, w_3), \theta) \approx$$



$$\hat{P}(X = (w_1, w_2, w_3))$$

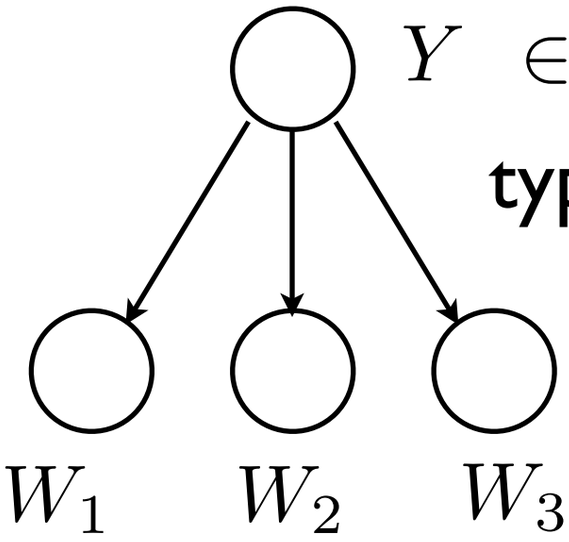
empirical distribution

	y	(w_1, w_2, w_3)
0.7	1	What is SVM?
0.3	0	What is SVM?
0.1	1	Mother's day flowers
0.9	0	Mother's day flowers
		...

M-step

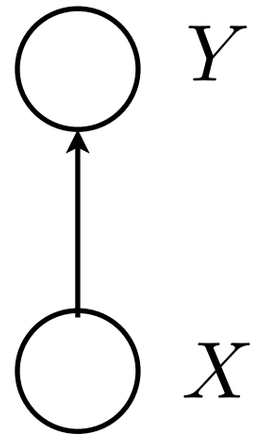
Mixture models: the EM algorithm

$$P(Y = y) = \theta(y)$$



$$\approx$$

$$P(Y = y | X = (w_1, w_2, w_3), \theta)$$



$$\hat{P}(X = (w_1, w_2, w_3))$$

empirical distribution

$$\prod_{i=1}^{3} P(W_i = w_i | Y = y) = \prod_{i=1}^{3} \theta_i(w_i | y)$$

$$\hat{\theta}_1(What|1) = \frac{0.7}{0.7 + 0.1}$$

$$\hat{\theta}_1(What|0) = \frac{0.3}{0.3 + 0.9}$$

...

	y	(w_1, w_2, w_3)
0.7	1	What is SVM?
0.3	0	What is SVM?
0.1	1	Mother's day flowers
0.9	0	Mother's day flowers
		...

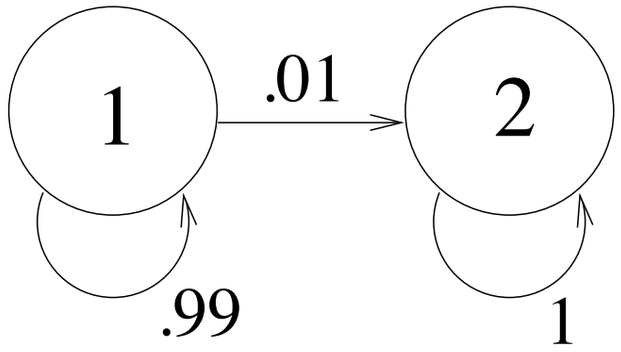
M-step

What do you need to know?

- Discriminative models and methods
  - linear classifiers, perceptron, max-margin hyperplane
  - non-linear classifiers, feature mappings, kernels
  - linear/non-linear regression
- Concepts
  - regularization, model selection, generalization
- Generative models and methods
  - mixture models, the EM-algorithm
  - hidden markov models
  - bayesian networks
- Decisions and actions
  - reinforcement learning

HMM example

$P(y_i|y_{i-1})$



$P(y_1)$

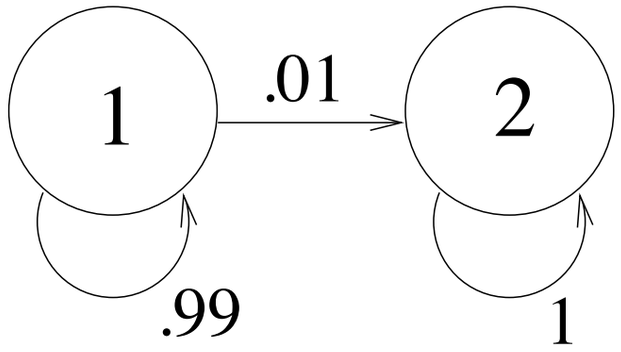
Y_1	
1	2
0.5	0.5

$P(x|y)$

	X			
Y	1	2	3	4
1	0	0.199	0.8	0.001
2	0.1	0	0.7	0.2

HMM example

$P(y_i|y_{i-1})$



$P(y_1)$

Y_1	
1	2
0.5	0.5

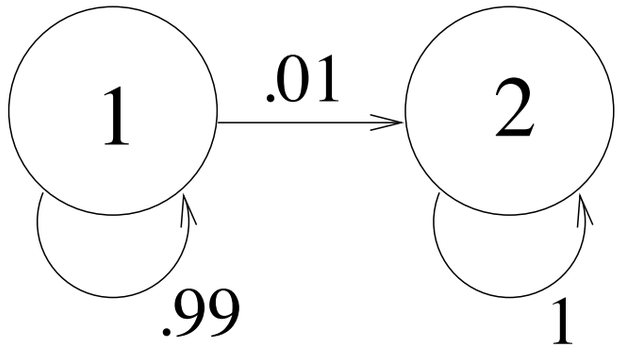
$P(x|y)$

Y	X 1	2	3	4
1	0	0.199	0.8	0.001
2	0.1	0	0.7	0.2

- What is an output sequence (of length 2) that cannot be generated from this HMM?

HMM example

$P(y_i|y_{i-1})$



$P(y_1)$

Y_1	
1	2
0.5	0.5

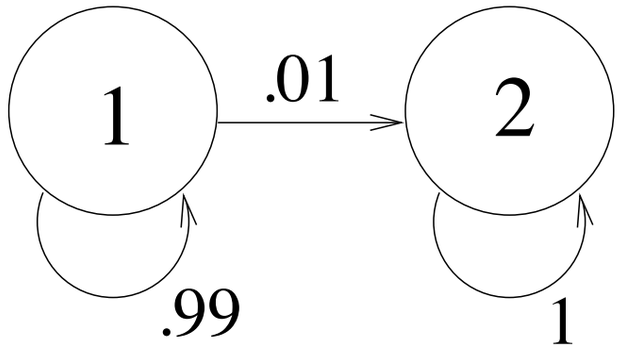
$P(x|y)$

Y	X 1	2	3	4
1	0	0.199	0.8	0.001
2	0.1	0	0.7	0.2

- We observe a sequence of length 6,064 and the last observation is x=3. What is the most likely hidden state at that point?

HMM example

$P(y_i|y_{i-1})$



$P(y_1)$

Y_1	
1	2
0.5	0.5

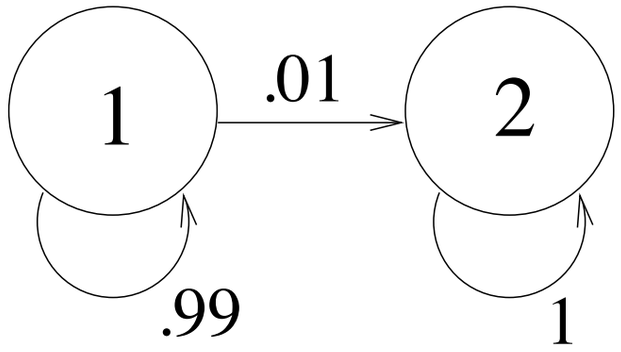
$P(x|y)$

Y	X 1	2	3	4
1	0	0.199	0.8	0.001
2	0.1	0	0.7	0.2

- Suppose we observe {3,3}. What is the most likely hidden state sequence?

HMM example

$P(y_i|y_{i-1})$



$P(y_1)$

Y_1	
1	2
0.5	0.5

$P(x|y)$

Y	X 1	2	3	4
1	0	0.199	0.8	0.001
2	0.1	0	0.7	0.2

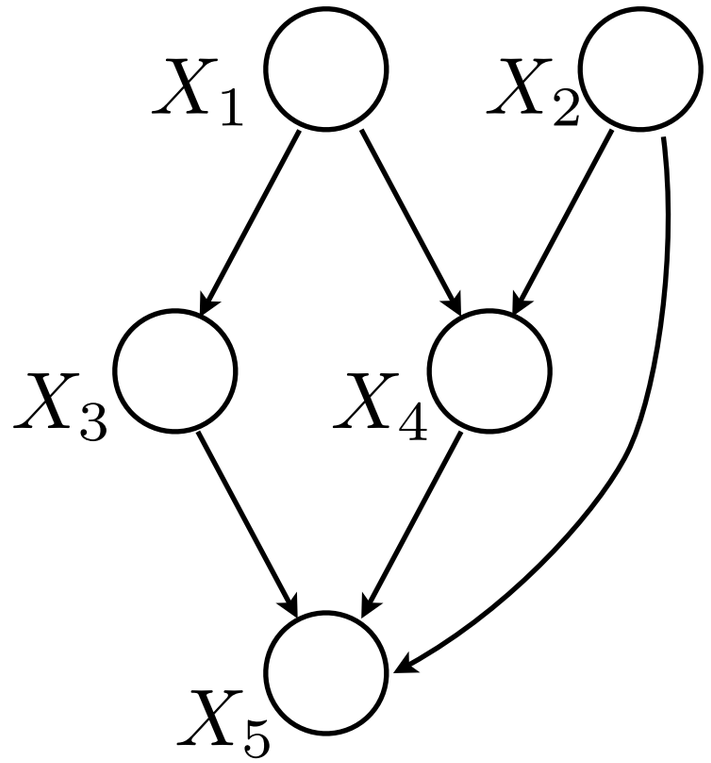
- Suppose we observe {3,3}. What is the most likely hidden state sequence?
- Suppose we observe {3,3,4}. What is the most likely hidden state sequence?

What do you need to know?

- Discriminative models and methods
  - linear classifiers, perceptron, max-margin hyperplane
  - non-linear classifiers, feature mappings, kernels
  - linear/non-linear regression
- Concepts
  - regularization, model selection, generalization
- Generative models and methods
  - mixture models, the EM-algorithm
  - hidden markov models
  - bayesian networks
- Decisions and actions
  - reinforcement learning

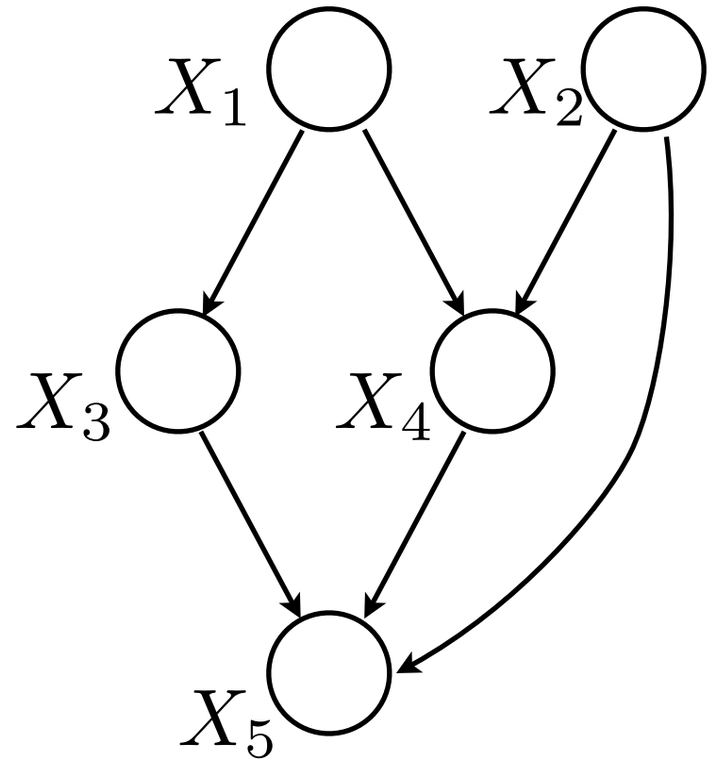
Bayesian network example

$X_i \in \{0, 1\}$



Bayesian network example

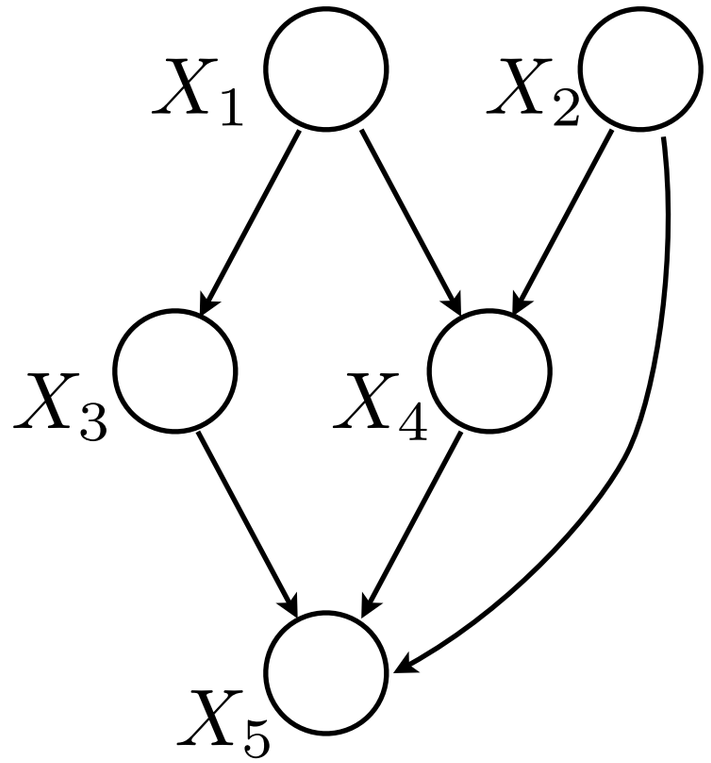
$X_i \in \{0, 1\}$



- How does the joint distribution factor over these variables?

Bayesian network example

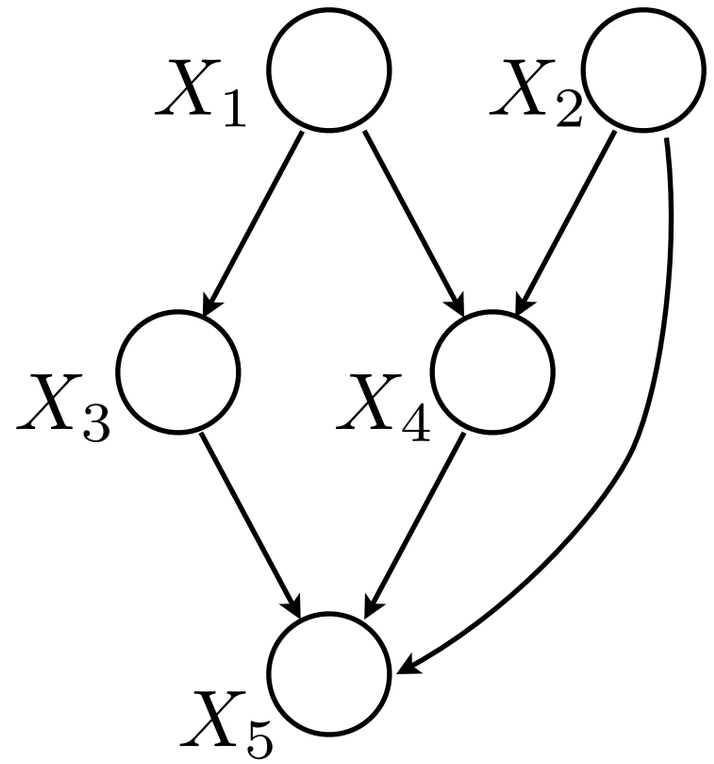
$X_i \in \{0, 1\}$



- How does the joint distribution factor over these variables?
- How many parameters are needed to specify the joint distribution?

Bayesian network example

$X_i \in \{0, 1\}$



- Which of the following statements hold for the variables in the graph (without any additional assumptions)?

( ) X_1 independent of X_2

( ) X_3 independent of X_4 given X_1

( ) knowing X_3 doesn't help predict X_2

( ) if we know X_4, knowing X_3 doesn't help predict X_2

Follow up courses

- 6.047 Computational Biology: Genomes, Networks, Evolution (U,G)
- 6.804J Computational Cognitive Science (U,G)
- 6.802 Computational Systems Biology (U,G)
- 6.867 Machine learning (G)
- 6.438 Algorithms for Inference (G)
- 6.864 Advanced Natural Language Processing (G)
- 6.869 Advances in Computer Vision (G)
- 9.520 Statistical Learning Theory and Applications (G)
- 6.???? Planning algorithms
- etc.

Follow up courses

- 6.047 Computational Biology: Genomes, Networks, Evolution (U,G)
  - Covers the algorithmic and machine learning foundations of computational biology, combining theory with practice. Principles of algorithm design, influential problems and techniques, and analysis of large-scale biological datasets. Topics include (a) genomes: sequence analysis, gene finding, RNA folding, genome alignment and assembly, database search; (b) networks: gene expression analysis, regulatory motifs, biological network analysis; (c) evolution: comparative genomics, phylogenetics, genome duplication, genome rearrangements, evolutionary theory. These are coupled with fundamental algorithmic techniques including: dynamic programming, hashing, Gibbs sampling, expectation maximization, hidden Markov models, stochastic context-free grammars, graph clustering, dimensionality reduction, Bayesian networks.

Follow up courses

- 6.804J Computational Cognitive Science (U,G)
  - Introduction to computational theories of human cognition. Focus on principles of inductive learning and inference, and the representation of knowledge. Computational frameworks covered include Bayesian and hierarchical Bayesian models; probabilistic graphical models; nonparametric statistical models and the Bayesian Occam's razor; sampling algorithms for approximate learning and inference; and probabilistic models defined over structured representations such as first-order logic, grammars, or relational schemas. Applications to understanding core aspects of cognition, such as concept learning and categorization, causal reasoning, theory formation, language acquisition, and social inference. Graduate students complete a final project.

Follow up courses

- 6.802 Computational Systems Biology (U,G)
  - Presents computational approaches and algorithms for contemporary problems in systems biology, with a focus on models of biological systems, including regulatory network discovery and validation. Topics include genotypes, regulatory factor binding and motif discovery, and whole genome RNA expression; regulatory networks (discovery, validation, data integration, protein-protein interactions, signaling, whole genome chromatin immunoprecipitation analysis); and experimental design (model validation, interpretation of interventions). Discusses computational methods, including directed and undirected graphical models, such as Bayesian networks, factor graphs, Dirichlet processes, and topic models. Multidisciplinary team-oriented final research project. Students taking graduate version complete additional assignments.

Follow up courses

- 6.867 Machine learning (G)
  - Principles, techniques, and algorithms in machine learning from the point of view of statistical inference; representation, generalization, and model selection; and methods such as linear/additive models, active learning, boosting, support vector machines, hidden Markov models, and Bayesian networks.

Follow up courses

- 6.438 Algorithms for Inference (G)
  - Introduction to statistical inference with probabilistic graphical models. Covers directed and undirected graphical models, factor graphs, and Gaussian models; hidden Markov models, linear dynamical systems.; sum−product and junction tree algorithms; forward−backward algorithm, Kalman filtering and smoothing; and min−sum algorithm and Viterbi algorithm. Presents variational methods, mean−field theory, and loopy belief propagation; and particle methods and filtering. Includes building graphical models from data; parameter estimation, Baum-Welch algorithm; structure learning; and selected special topics.

Follow up courses

- 6.864 Advanced Natural Language Processing (G)
  - Graduate introduction to natural language processing, the study of human language from a computational perspective. Syntactic, semantic and discourse processing models. Emphasis on machine learning or corpus-based methods and algorithms. Use of these methods and models in applications including syntactic parsing, information extraction, statistical machine translation, dialogue systems, and summarization.

Follow up courses

- 6.869 Advances in Computer Vision (G)
  - Advanced topics in computer vision with a focus on the use of machine learning techniques and applications in graphics and human-computer interface. Topics include image representations, texture models, structure-from-motion algorithms, Bayesian techniques, object and scene recognition, tracking, shape modeling, and image databases. Applications may include face recognition, multimodal interaction, interactive systems, cinematic special effects, and photorealistic rendering. Covers topics complementary to 6.801/6.866; these subjects may be taken in sequence.

Follow up courses

- 9.520 Statistical Learning Theory and Applications (G)
  - Focuses on the problem of supervised and unsupervised learning from the perspective of modern statistical learning theory, starting with the theory of multivariate function approximation from sparse data. Develops basic tools such as regularization, including support vector machines for regression and classification. Derives generalization bounds using stability. Discusses current research topics such as manifold regularization, sparsity, feature selection, bayesian connections and techniques. Discusses applications in areas such as computer vision, speech recognition, and bioinformatics. Also covers advances in the neuroscience of the cortex and their impact on learning theory and applications. Includes a final project.

Follow up courses

- 6.???? Planning algorithms
  - Introduction to algorithms for planning action sequences with applications in artificial intelligence, robotics and computer games. The course covers a broad spectrum of representations and algorithms from (a) symbolic planning, (b) robot motion planning and (c) probabilistic planning . Topics include: state-space search, heuristics, STRIPS planning, configuration-space representation, sampling-based motion planning, decision theory, Markov decision processes and partially observable Markov decision processes.

Follow up courses

- 6.047 Computational Biology: Genomes, Networks, Evolution (U,G)
- 6.804J Computational Cognitive Science (U,G)
- 6.802 Computational Systems Biology (U,G)
- 6.867 Machine learning (G)
- 6.438 Algorithms for Inference (G)
- 6.864 Advanced Natural Language Processing (G)
- 6.869 Advances in Computer Vision (G)
- 9.520 Statistical Learning Theory and Applications (G)
- 6.???? Planning algorithms
- etc.

The end