CS 3750 Machine Learning

Lecture 2

Advanced Machine Learning

Milos Hauskrecht
milos@cs.pitt.edu
5329 Sennott Square, x4-8845

http://www.cs.pitt.edu/~milos/courses/cs3750/

Tentative topics

- **Review:** supervised learning, density estimation
- **Extending standard learning framework:**
  - **sparsity, learning to rank, multiple task**
- **Low dimensional representation of data**
  - **Component analysis and their applications**
    - PCA, LSA, PLSA, pPCA, ICA, etc
  - **Latent variable models**
    - Variational approximations
- **Kernels**
  - Kernel methods, Kernel-PCA, string kernels, etc.
- **Non-parametric models and methods:**
  - Graph-based kernels for classification and clustering
  - Metric learning
  - Gaussian processes

Learning

Starts with data & prior knowledge

Typical steps in learning:
- Define a model space
- Define an objective criterion: criterion for measuring the goodness of a model (fit to data)
- Optimization: finding the best model

Alternative: optimization is replaced with the inference, e.g. Bayesian inference in the Bayesian learning

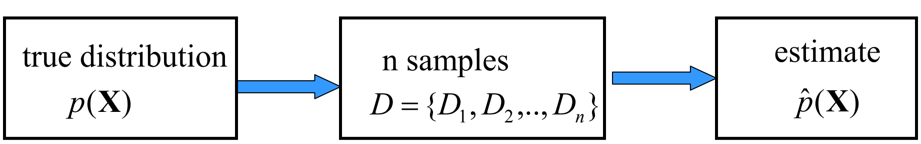
Evaluation/application:
- Model learned from the training data
- generalization to the future (test) data

Density estimation

Data: $D = \{D_1, D_2, .., D_n\}$

$D_i = \mathbf{x}_i$ a vector of attribute values

Objective: try to estimate the underlying true probability distribution over variables $\mathbf{X}$, $p(\mathbf{X})$, using examples in D



Standard (iid) assumptions: Samples
- **are independent of each other**
- **come from the same (identical) distribution (fixed $p(\mathbf{X})$)**

Density estimation

Types of density estimation:

Parametric

- the distribution is modeled using a set of parameters Θ

$$p(\mathbf{X} \mid \Theta)$$

- **Example:** mean and covariances of multivariate normal
- **Estimation:** find parameters $\hat{\Theta}$ that fit the data D the best

Non-parametric

- The model of the distribution utilizes all examples in D
- As if all examples were parameters of the distribution
- The density for a point x is influenced by examples in its neighborhood

Basic criteria

What is the best set of parameters?

- **Maximum likelihood (ML)**

maximize $p(D \mid \Theta, \xi)$

ξ - represents prior (background) knowledge

- **Maximum a posteriori probability (MAP)**

maximize $p(\Theta \mid D, \xi)$

Selects the mode of the posterior

$$p(\Theta \mid D, \xi) = \frac{p(D \mid \Theta, \xi)\, p(\Theta \mid \xi)}{p(D \mid \xi)}$$

Example. Bernoulli distribution.

Outcomes: two possible values – 0 or 1 (head or tail)
Data: D a sequence of outcomes x_i with 0,1 values

Model: probability of an outcome 1 θ
probability of 0 $(1-\theta)$

$$P(x_i \mid \theta) = \theta^{x_i}(1-\theta)^{(1-x_i)}$$ **Bernoulli distribution**

Objective:

We would like to estimate the probability of seeing 1:

$\hat{\theta}$

Maximum likelihood (ML) estimate.

Likelihood of data:

$$P(D \mid \theta, \xi) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{(1-x_i)}$$

Maximum likelihood estimate

$$\theta_{ML} = \arg\max_{\theta} P(D \mid \theta, \xi)$$

Optimize log-likelihood

$$l(D,\theta) = \log P(D \mid \theta, \xi) = \log \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{(1-x_i)} =$$

$$\sum_{i=1}^{n} x_i \log\theta + (1-x_i)\log(1-\theta) = \log\theta \underline{\sum_{i=1}^{n} x_i} + \log(1-\theta) \underline{\sum_{i=1}^{n}(1-x_i)}$$

N_1 - number of 1s seen N_2 - number of 0s seen

Maximum likelihood (ML) estimate.

Optimize log-likelihood

$$l(D,\theta) = N_1 \log\theta + N_2 \log(1-\theta)$$

Set derivative to zero

$$\frac{\partial l(D,\theta)}{\partial\theta} = \frac{N_1}{\theta} - \frac{N_2}{(1-\theta)} = 0$$

Solving $\theta = \frac{N_1}{N_1 + N_2}$

ML Solution: $\theta_{ML} = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2}$

Maximum a posteriori estimate

Maximum a posteriori estimate

– Selects the mode of the posterior distribution

$$\theta_{MAP} = \arg\max_{\theta} p(\theta \mid D, \xi)$$

$$p(\theta \mid D, \xi) = \frac{P(D \mid \theta, \xi) p(\theta \mid \xi)}{P(D \mid \xi)}$$ **(via Bayes rule)**

$P(D \mid \theta, \xi)$ - is the likelihood of data

$$P(D \mid \theta, \xi) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{(1-x_i)} = \theta^{N_1} (1-\theta)^{N_2}$$

$p(\theta \mid \xi)$ - is the prior probability on θ

How to choose the prior probability?

Prior distribution

Choice of prior: Beta distribution

$$p(\theta \mid \xi) = Beta(\theta \mid \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_2 - 1}$$

Why?

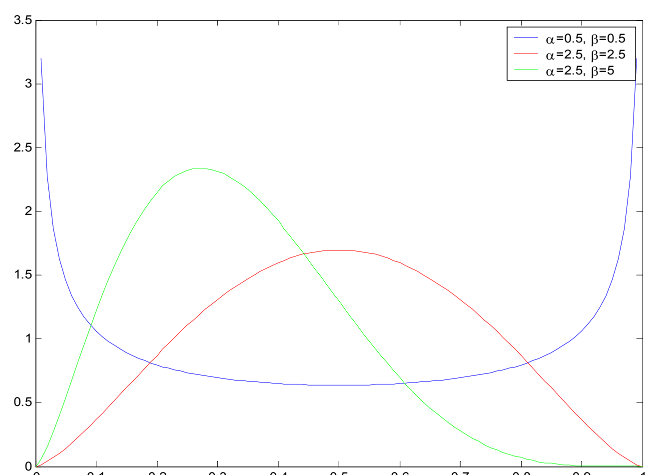
Beta distribution "**fits**" binomial sampling - **conjugate choices**

$$P(D \mid \theta, \xi) = \theta^{N_1}(1-\theta)^{N_2}$$

$$p(\theta \mid D, \xi) = \frac{P(D \mid \theta, \xi) Beta(\theta \mid \alpha_1, \alpha_2)}{P(D \mid \xi)} = Beta(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2)$$

MAP Solution: $$\theta_{MAP} = \frac{\alpha_1 + N_1 - 1}{\alpha_1 + \alpha_2 + N_1 + N_2 - 2}$$

Beta distribution



Bayesian learning

- **Both ML or MAP pick one parameter value**
  - Is it always the best solution?
- **Full Bayesian approach**
  - Remedies the limitation of one choice
  - Keeps and uses a complete posterior distribution
- **How is it used? Assume we want:** $P(\Delta \mid D, \xi)$
  - Considers all parameter settings and averages the result

$$P(\Delta \mid D, \xi) = \int_{\theta} P(\Delta \mid \theta, \xi) p(\theta \mid D, \xi) d\theta$$

  - **Example:** predict the result of the next outcome
    - Choose outcome 1 if $P(x=1 \mid D, \xi)$ is higher

Modeling complex multivariate distributions

How to model complex multivariate distributions $\hat{p}(\mathbf{X})$ with large number of variables?

One solution:

- **Decompose the distribution. Reduce the number of parameters, using some form of independence.**

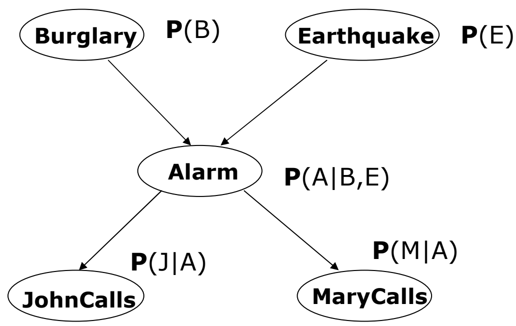
Two models:

- **Bayesian belief networks (BBNs)**
- **Markov Random Fields (MRFs)**

- **Learning.** Relies on the decomposition.

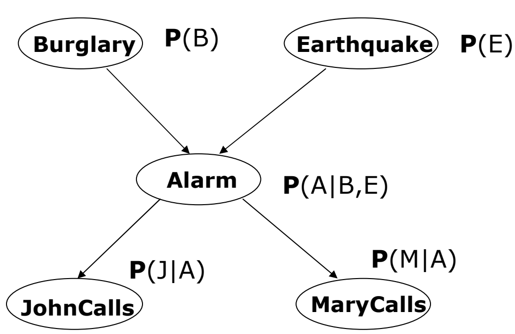
Bayesian belief network.

1. **Directed acyclic graph**
   - **Nodes** = random variables
   - **Links** = direct (causal) dependencies between variables
     - Missing links encode independences

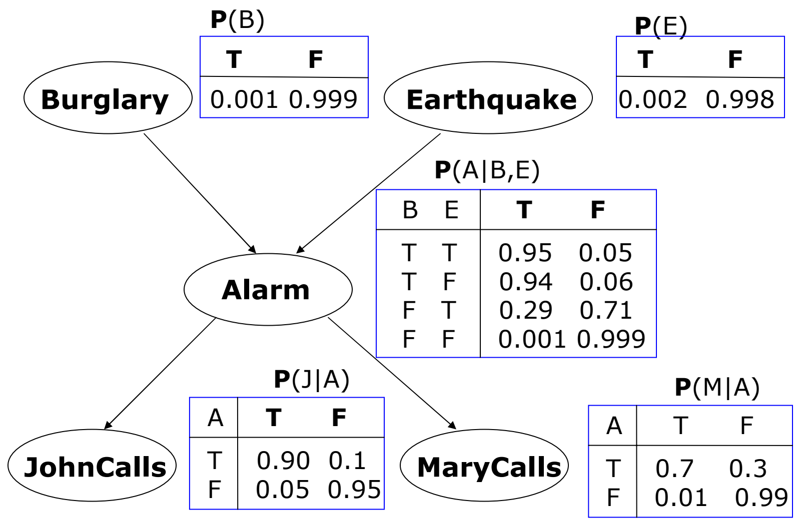


Bayesian belief network.

2. **Local conditional distributions**
   - relate variables and their parents



Bayesian belief network.

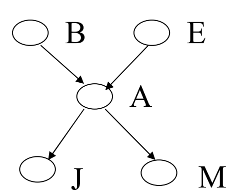


Full joint distribution in BBNs

Full joint distribution is defined in terms of local conditional distributions (obtained via the chain rule):

$$\mathbf{P}(X_1, X_2, .., X_n) = \prod_{i=1,..n} \mathbf{P}(X_i \mid pa(X_i))$$

Example:



Assume the following assignment of values to random variables

$B=T, E=T, A=T, J=T, M=F$

Then its probability is:

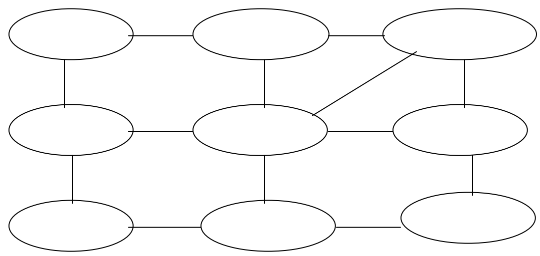
$$P(B=T, E=T, A=T, J=T, M=F) =$$
$$P(B=T)P(E=T)P(A=T \mid B=T, E=T)P(J=T \mid A=T)P(M=F \mid A=T)$$

Markov Random Fields (MRFs)

Undirected graph

- **Nodes** = random variables
- **Links** = direct relations between variables

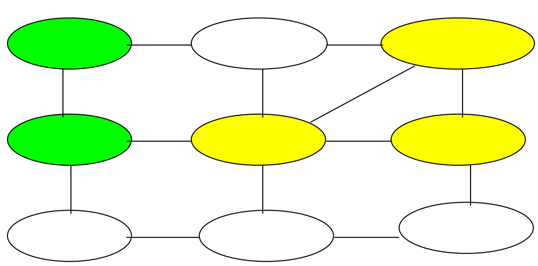
- BBNs used to model **asymetric** dependencies (most often causal),
- MRFs model **symmetric** dependencies (bidirectional effects) such as spatial dependences



Markov Random Fields (MRFs)

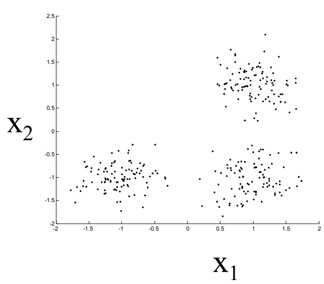
A probability distribution is defined in terms of potential functions defined over cliques of the graph

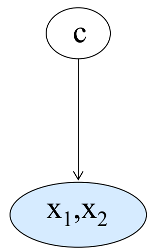
$$\mathbf{P}(X_1, X_2, .., X_n) = \frac{1}{Z} \prod_{C_i \in cliques(G)} \Psi(C_i)$$

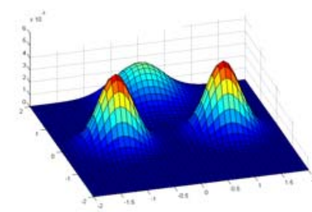


Latent variable models

- We can have a model with hidden variables
- Hidden variables may help us to induce the decomposition of a complex distribution

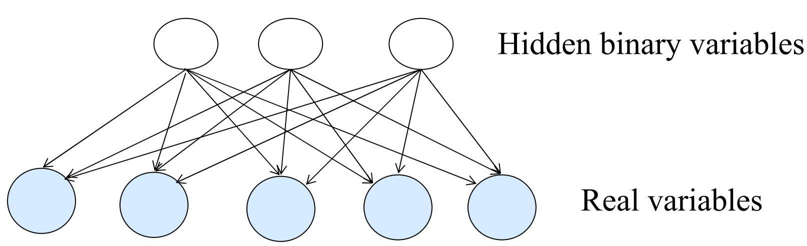






Latent variable models

- More general latent variable models
- Various relations in between hidden and observable variables
- **Example:** Continuous vector quantizer (CVQ) model



- **Possible uses:**
- A probabilistic model
- A low dimensional representation of observable data

Copula distributions

- Copula defines a joint distribution function for random variables U1,U2, . .,Uk each of which is marginally uniformly distributed on (0, 1).
- **Important (Sklar's theorem):** A distribution function for a multivariate X can be written as a copula of marginal distribution functions
- Copula is used to model all dependences in between components of X