Directed Graphical Models

We could formulate and solve complicated probabilistic models purely by algebraic manipulation. However, probabilistic graphical models offer

1. Simple way to **visualize structure** of a probabilistic model

2. Simple analysis of model properties, e.g. **conditional independence** properties.

3. Complex computations, required for **inference and learning**, can be expressed in terms of simple graphical manipulations.
D-separation

The d-separation criterion ascertains whether a particular conditional independence statement $A \perp B \mid C$ is implied by a given directed acyclic graph, where $A$, $B$ and $C$ are arbitrary non-intersecting sets of nodes.

If all paths between $A$ and $B$ are blocked, then $A$ is d-separated from $B$ by $C$.

A path is blocked if it includes a node such that either

1) the arrows on the path are head-to-tail or tail-to-tail and the node is in the set $C$.

2) the arrows on the path meet head-to-head and neither the node nor its descendants is in the set $C$. 
D-separation

A path is blocked if it includes a node such that either

1) the arrows on the path are head-to-tail or tail-to-tail and the node is in the set $C$.

$$P(A, B, C) = P(A|C) P(B|C) P(C)$$  
$$P(A, B, C) = P(B|C) P(C|A) P(A)$$
A path is blocked if it includes a node such that either

2) the arrows on the path meet head-to-head and neither the node nor its descendants is in the set $C$.

$$P(A, B, C) = P(C | A, B) P(A) P(B)$$

$$P(A, B, C, D) = P(D | C) P(C | A, B) P(A) P(B)$$
Example: d-separation

(a) Path from $A$ to $B$ is neither blocked by $F$ nor by $E$.

(b) Path from $A$ to $B$ is blocked by $F$ and by $E$. 

Assignment 6
The Markov blanket of a node $x$ comprises the set of parents, children and co-parents of the node. It has the property that the conditional distribution of $x$, conditioned on all the remaining variables in the graph, is dependent only on the variables in the Markov blanket.

In order to apply the d-separation criterion consider all possible types of paths from the central node $x$ to all possible nodes external to the Markov blanket.
Naive Bayes Classifiers

Key assumption: Conditional independence

$$P(x_1, x_2, \ldots, x_d | c_i) = \prod_{i=1}^{d} P(x_i | c_i)$$

An advantage of the Naive Bayes classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification.

Target functions have to take on values from a finite set of classes.

For the estimation of the model parameters non-discrete features need to be discretized first.
Consider the case of binary variables, i.e. $x_i$ and $c \in \{0,1\}$.

In many practical applications, parameter estimation for naive Bayes models uses the method of \textbf{max. likelihood} (ML). In case of $N$ i.i.d. samples the likelihood is a Bernoulli distribution

$$\text{Bern}(x|\mu) = \mu^x (1-\mu)^{1-x}; \quad P(x=1|\mu) = \mu; \quad P(x=0|\mu) = 1-\mu;$$

$$P(x|\mu) = \mu^m (1-\mu)^l, \quad m = \sum_{i=1}^{N} x_i, \quad l = N-m$$

with the ML solution

$$\frac{d}{d\mu} \log P(x|\mu) = \frac{d}{d\mu} \left( m \log \mu + (N-m) \log (1-\mu) \right) = 0 \quad \rightarrow \quad \mu_{ML} = \frac{m}{N}$$

Problem with small sample sizes, for which $\mu$ is either 0 or 1.
MAP solution

Take a Bayesian approach and estimate MAP solution for a particular prior distribution:

**Conjugate priors:** Posterior has the same functional form as the prior.

\[
\text{Beta}(\mu | a, b) \propto \mu^{a-1}(1-\mu)^{b-1}, \quad a, b \in \mathbb{R}^+
\]

\[
P(\mu | x, a, b) = P(x | \mu, a, b) \cdot P(\mu | a, b) \propto \mu^{m+a-1}(1-\mu)^{N-m+b-1}
\]

\[
\frac{d}{d\mu} \log P(\mu | x, a, b) = 0 \quad \Rightarrow \quad \mu_{MAP} = \frac{m+a-1}{N+a+b-2}
\]

For \(a = b = 2\) we get

\[
\mu_{MAP} = \frac{m+1}{N+2}
\]
This is equivalent to the ML solution that is usually used + correction.

If a given class and feature value never occur together in the training set then the frequency-based probability estimate will be zero. This will wipe out all information in the other probabilities when they are multiplied.

It is therefore often desirable to incorporate a small-sample correction in all probability estimates such as the Laplace estimate:

\[
P(x_i = v_j | c_k) = \frac{n_{ijk} + 1}{n_k + s_i}
\]

- \(n_{ijk}\): number of examples in class \(c_k\) where \(x_i = v_j\).
- \(n_k\): number of samples in \(c_k\).
- \(s_i\): number of possible values for \(x_i\).
Example: Spam Dataset

Creators: Hewlett-Packard Labs, Palo Alto, CA
Donor: George Forman (c) Generated: June-July 1999

Determine whether a given email is spam or not (~7% misclass. error).

False positives (marking good mail as spam) are very undesirable. For zero false positives in the training/testing set 20-25% of the spam passed through the filter.

The original $d=57$ real-valued features are converted to binary random variable by making all values less than or equal to the median

48 real = percentage of words in the e-mail that match {money, credit, georg, address,...},
6 real = percentage of characters in the e-mail that match {;, [, ], !, $, #},
1 real = average length of uninterrupted sequences of capital letters
1 integer = length of longest uninterrupted sequence of capital letters
1 integer = total number of capital letters in the e-mail
Learn a Naive Bayes Classifier

1. Separate the examples into their respective classes using the class labels provided.

2. Calculate the marginal probability for the class labels

   \[ P(c_i) = \frac{N_i}{N} \]

   where \( N_1 \) and \( N_2 \) are the number of examples that have class values 1 and 2 (respectively), and let \( N \) be the total number of examples.

3. For each class value \( c_k \), for each attribute \( x_i \), and for each of the possible value \( j \) of the attribute \( x_i \), calculate the Laplace estimate

   \[ P(x_i = j | c_k) = \frac{n_{ijk} + 1}{n_k + 2}, \quad j \in \{1,2\} \]

4. Ensure that the conditional probabilities sum to 1.
Using a Naive Bayes Classifier for Classification

1. Find the most likely class label, i.e. the class $c_i$ for which the probability

$$p(c_i | x_1, \ldots, x_d) = \frac{p(c_i) \prod_{j=1}^{d} p(x_j | c_i)}{p(x_1, \ldots, x_d)}$$

is maximum.

2. To find the maximum of this expression (over $c_i$) we only need to calculate the numerators and not the denominator.

3. Do this computation in log-space to get around the underflow problem.

4. To predict a class select $c_i$ for which the posterior is largest.
To apply the sum-product algorithm we introduce a new graphical construction called **Factor Graphs**.

\[
P(x_1, x_2, x_3) = P(x_1)P(x_2)P(x_3|x_1, x_2)
\]

\[
P(x) = \prod_s f_s(x_s)
\]

\[
f_1(x_1, x_2, x_3) = P(x_1)P(x_2)P(x_3|x_1, x_2)
\]

\[
f_1(x_1) = P(x_1)
\]

\[
f_2(x_2) = P(x_2)
\]

\[
f_3(x_1, x_2, x_3) = P(x_3|x_1, x_2)
\]
Sum-Product Algorithm

Used for exact inference, i.e. to evaluate marginals over nodes.

\[
\mu_{f_s \rightarrow x}(x) = \sum_{x_s} f_s(x_s) \prod_{m \in \text{ne}(f_s) \setminus x} \mu_{x_m \rightarrow f_s}(x_m)
\]

\[
\mu_{x_m \rightarrow f_s}(x_m) = \prod_{l \in \text{ne}(x_m) \setminus f_s} \mu_{f_l \rightarrow x_m}(x_m)
\]

Initialization of leaves:

\[\mu_{x \rightarrow f}(x) = 1\]

\[\mu_{f \rightarrow x}(x) = f(x)\]
Sum-Product Algorithm

How to calculate marginal distributions:

\[ P(x) = \prod_{s \in ne(x)} \mu_{f_s \rightarrow x}(x) \]
\[ P(x_s) = f_s(x_s) \prod_{i \in ne(f_s)} \mu_{x_i \rightarrow f_s}(x_i) \]

If marginals of only a few \( x_i \) of the \( x_s \) associated with a factor \( f_s \) are required then sum out the remaining variables, e.g.

\[ P(x_1, x_2, x_3) = f_s(x_1, x_2, x_3) \mu_{x_1 \rightarrow f_s}(x_1) \mu_{x_2 \rightarrow f_s}(x_2) \mu_{x_3 \rightarrow f_s}(x_3) \]
\[ P(x_1, x_2) = \sum_{x_3} P(x_1, x_2, x_3) \]
Fred lives in Los Angeles and commutes 60 miles to work. Whilst at work, he receives a phone-call from his neighbor saying that Fred's burglar alarm is ringing. What is the probability that there was a burglar in his house today? While driving home to investigate, Fred hears on the radio that there was a small earthquake that day near his home. `Oh', he says, feeling relieved, `it was probably the earthquake that set off the alarm'. What is the probability that there was a burglar in his house? (After Pearl, 1988, see book of MacKay).

\[ P(b, e, a, p, r) = P(b)P(e)P(a \mid b, e)P(p \mid a)P(r \mid e) \]
Example: Burglar Alarm

\[
\begin{align*}
\mu_{f_2 \rightarrow b} &= P(b) \\
\mu_{f_1 \rightarrow e} &= P(e) \\
\mu_{r \rightarrow f_3} &= 1 \\
\mu_{f_3 \rightarrow e} &= \sum_r P(r|e) \mu_{r \rightarrow f_3} \\
\mu_{e \rightarrow f_4} &= \mu_{f_1 \rightarrow e} \cdot \mu_{f_3 \rightarrow e} \\
\mu_{p \rightarrow f_5} &= 1 \\
\mu_{f_5 \rightarrow a} &= \sum_p P(p|a) \mu_{p \rightarrow f_5} \\
\mu_{a \rightarrow f_4} &= \mu_{f_5 \rightarrow a} \\
\mu_{f_4 \rightarrow b} &= \sum_{e,a} P(a|e,b) \mu_{a \rightarrow f_4} \cdot \mu_{e \rightarrow f_4} \\
\end{align*}
\]

\[P(b) = \mu_{f_2 \rightarrow b} \cdot \mu_{f_4 \rightarrow b}\]
Example: Burglar Alarm

1. Estimate each of the prior and conditional probabilities from the data using the ML solution for the Bernoulli distribution.

2. Clamp the conditional or prior distribution for each observed variable to the corresponding value $a_o$.

   \[ P(a|e,b) = P(a|e,b) \delta(a-a_o), \text{ i.e. only at } a_o \text{ non-zero.} \]

   \[ \mu_{f_4 \rightarrow b} = \sum_{e,a} P(a|e,b) \mu_{a \rightarrow f_4} \cdot \mu_{e \rightarrow f_4} \]

3. Take care of the proper normalization, that is maybe changed due to clamping.
Lung Cancer Example

TABLE 2.1
Preliminary choices of nodes and values for the lung cancer example

<table>
<thead>
<tr>
<th>Node name</th>
<th>Type</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pollution</td>
<td>Binary</td>
<td>{low, high}</td>
</tr>
<tr>
<td>Smoker</td>
<td>Boolean</td>
<td>{T, F}</td>
</tr>
<tr>
<td>Cancer</td>
<td>Boolean</td>
<td>{T, F}</td>
</tr>
<tr>
<td>Dyspnoea</td>
<td>Boolean</td>
<td>{T, F}</td>
</tr>
<tr>
<td>X-ray</td>
<td>Binary</td>
<td>{pos, neg}</td>
</tr>
</tbody>
</table>