Topic 8: GRAPHICAL MODELS

STAT 37710/CAAM 37710/CMSC 35400 Machine Learning
Risi Kondor, The University of Chicago

Three types of “Probability”

1. **Frequency of repeated trials:** if an experiment is repeated infinitiely many times, $0 \leq p(A) \leq 1$ is the fraction of times that the outcome will be A. Typical example: number of times that a coin comes up heads. $\rightarrow$ Frequentist probability.
2. **Degree of belief:** A quantity obeying the same laws as the above, describing how likely we think a (possibly deterministic) event is. Typical example: the probability that the Earth will warm by more than 5° F by 2100. $\rightarrow$ Bayesian probability.
3. **Subjective probability:** “I’m 110% sure that I’ll go out to dinner with you tonight.”

Mixing these three notions is a source of lots of trouble. We will start with the frequentist interpretation and then discuss the Bayesian one.

Why do we need probability for ML?

Two distinct reasons:

1. To analyze, understand and predict the performance of learning algorithms (Statistical Learning Theory, PAC model, etc.)
2. To build flexible and intuitive **probabilistic models.**

Probabilistic vs. Algorithmic learning

- Algorithmic ML (e.g., SVMs):
  - Strictly focus on the task at hand $\rightarrow$ discriminative
  - Black box
  - Algorithms often motivated directly by optimization methods $\rightarrow$ fast
  - Examples: the perceptron, SVM, etc.
  - “Frequentist”
- Probabilistic ML (e.g., graphical models):
  - Everything in the world is a random variable $\rightarrow$ generative
  - Flexible modeling framework for incorporating prior knowledge
  - Models are often expressed with graphs $\rightarrow$ efficient message passing algorithms
  - Example: k–means clustering
  - “Bayesian”

[Breiman: Statistical modeling: the two cultures]

Joint probabilities and independence

Machine learning applications often involve a large number of variables (features) $X_1, \ldots, X_n$.

- The **conditional probability** of X_i given X_j is

$$p(x_i|x_j) = \mathbb{P}(X_i = x_i \mid X_j = x_j) \qquad p(x_i, x_j) = p(x_i|x_j)\, p(x_j).$$

- X_i and X_j are **independent** (denoted $X_i \perp\!\!\!\perp X_j$) if

$$p(x_i|x_j) \text{ is indep of } x_j \qquad \Longleftrightarrow \qquad p(x_i, x_j) = p(x_i)\, p(x_j).$$

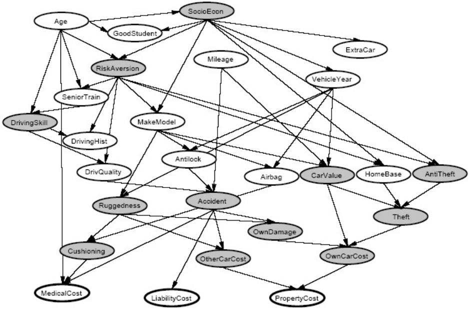
- X_i is **conditionally independent** of X_j given X_k (denoted $X_i \perp\!\!\!\perp X_j | X_k$) if

$$p(x_i, x_j | x_k) = p(x_i|x_k)\, p(x_j|x_k).$$

IDEA: When faced with a large number of features, use our prior knowledge of indepdencies to make learning easier.

Directed graphical models

Also called Bayes nets or Belief Networks. Each vertex $v \in V$ corresponds to a random variable. Graph must be acyclic but not necessarily a tree.



The general form of the joint distribution of all the variables is

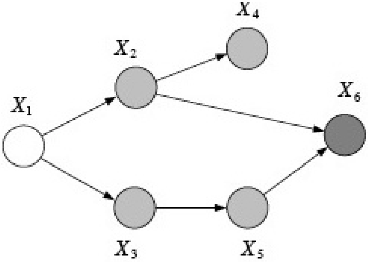
$$p(\mathbf{x}) = \prod_{v \in V} p(x_v | \mathbf{x}_{\mathsf{pa}(v)}),$$

where $\mathsf{pa}(v)$ are all the parents of v in the graph.

Directed graphical models

Assuming that $X_1, \ldots, X_6$ are binary random variables, how many numbers are need to describe their joint distribution? $2^6 - 1 = 63$.

Now what if we know that they conform to this Bayes net?



Each $p(x_i|x_j)$ corresponds to a 2×2 table, but rows sum to 1, so only 2 numbers required. $p(x_6|x_2, x_5)$ requires 4 numbers.
Total: $1 + 2 + 2 + 2 + 2 + 4 = 13$. Quite a saving!

Example: Markov chains

- If $x_1, x_2, \ldots$ is a series of (discrete or continuous) random variables corresponding to a process evolving in time, then x_t should only depend on what happened in the past:

$$p(x_t|x_1, \ldots, x_{t-1}, x_{t+1}, \ldots) = p(x_t|x_1, \ldots, x_{t-1}).$$

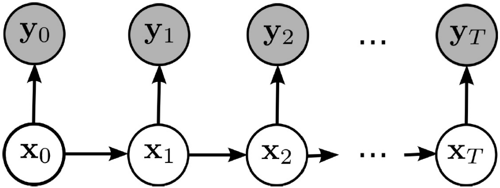
- The sequence $x_1, x_2, \ldots$ is said to be a $\mathbf{k}$**'th order Markov chain** if

$$p(x_t|x_1, \ldots, x_{t-1}, x_{t+1}, \ldots) = p(x_t|x_{t-1}, \ldots, x_{t-k}).$$

- A (first order) Markov chain is said to be **stationary** if the $p(x_t|x_{t-1})$ **transition probabilities** are independent of t,

$$p(x_t|x_{t-1}) = M_{x_t, x_{t-1}}.$$

Hidden Markov Models (HMM)



An HMM is a Markov chain of unobserved random variables $x_1, x_2, \ldots,$ each of which is related to an oberved random variable $y_1, y_2, \ldots$.

Example: Tracking, part of speech tagging, phonemes, physiological states of babies,...

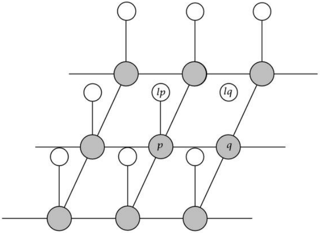
Applications of HMMs

HMMs and related state space models are widely applied in

- speech recognition (which phoneme/word/etc.)
- part of speech tagging (is it a NP, VP, etc.)?
- biological sequence analysis (intron or extron)?
- time series analysis (finance, climate, etc.)
- robotics (what is the actual location of the robot)?
- tracking

Undirected graphical models

Also called Markov Random Fields. Graph can be any undirected graph. Common example used for image segmentation:

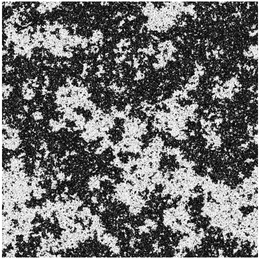


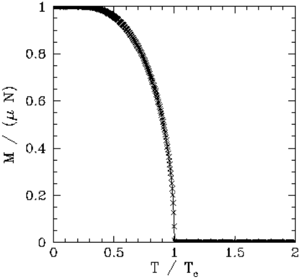
The general form of the joint distribution over all the variables is

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{c \in \text{Cliques}(\mathcal{G})} \phi_c(\mathbf{x}_c)$$

where each ϕ_c is a potentially different **clique potential** (just a positive function) and Z is the **normalizing factor** $Z = \sum_{\mathbf{x}} \prod_{c \in \text{Cliques}(\mathcal{G})} \phi_c(\mathbf{x}_c)$.

Example: the Ising model



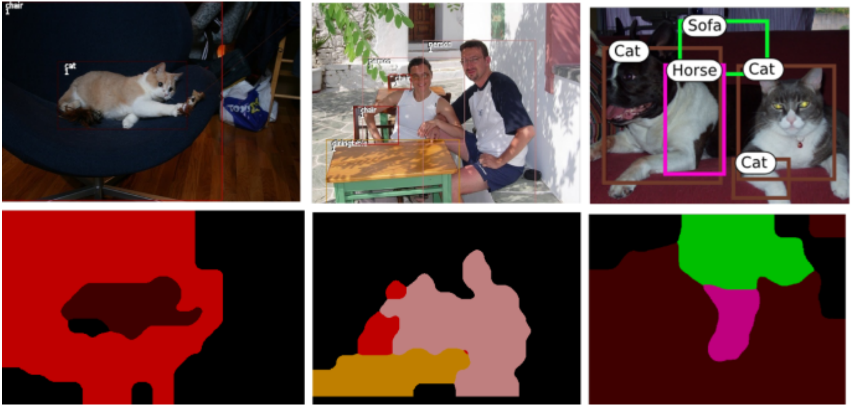


Imagine an infinite grid of $\{-1,+1\}$ valued random variables in which neighboring variables are connected by the potential

$$\phi(x_i, x_j) = e^{-\beta/2(x_i - x_j)^2}.$$

Simple model of ferromagnetism. Exhibits a **phase transition.**

Example: MRFs for segmentation



Purpose of graphical models

In ML we often have a large number of variables related in complicated ways.

Graphical models

- capture prior knowledge about relationships between variables
- provide a compact representation of distributions over many variables
- define a specific hypothesis class
- help with figuring out causality
- the variables can be either discrete (e.g., “airbag yes/no”), continous (e.g., “value”) or a mixture of both types

Tasks for graphical models

- Model selection (i.e., learn the graph itself from data)
- Learn the parameters of the model from data (i.e., the individual conditionals or clique potentials)
- Deduce conditional independence relations
- Infer marginals and conditional distributions

Inference

Partition V, the set of nodes, into three sets:

1. the set O of observed nodes
2. the set Q of query nodes
3. the set L of latent nodes

Interested in $$p(\mathbf{x}_Q|\mathbf{x}_O) = \frac{\sum_{\mathbf{x}_L} p(\mathbf{x}_Q, \mathbf{x}_L, \mathbf{x}_O)}{\sum_{\mathbf{x}_L, \mathbf{x}_Q} p(\mathbf{x}_Q, \mathbf{x}_L, \mathbf{x}_O)}$$

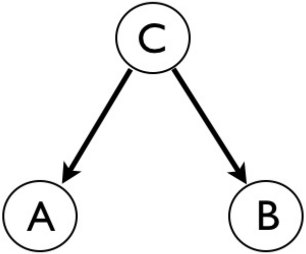
Essential for both

- **Training**, when we are trying to learn the distribution of some of the nodes from data.
- **Prediction**, when we are trying to predict the values of some nodes (the output) given the values of some other nodes (the input)

Question: How can we do this in less than $m^{|Q|+|L|}$ time?

Directed graphical models (Bayes nets)

Common cause

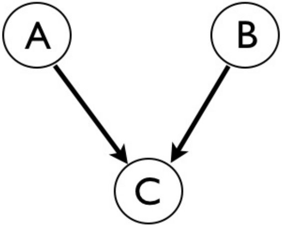


$$X_A \not\perp\!\!\!\perp X_B \quad \text{but} \quad X_A \perp\!\!\!\perp X_B \,|\, X_C$$

Therefore, if C is *observed*, then A and B become independent.

Example: Lung cancer $\perp\!\!\!\perp$ Yellow teeth $|$ Smoking

Explaining away



$$X_A \perp\!\!\!\perp X_B \qquad \text{but} \qquad X_A \not\!\perp\!\!\!\perp X_B \,|\, X_C$$

Therefore, if C is *not oberseved* (and neither are any of its descendents) then A and B become independent.

Example: Burglary $\not\!\perp\!\!\!\perp$ Earthquake $|$ Alarm

D–separation

Is X independent of Y given the set of nodes S?

An underected path from X to Y is said to be **blocked** if

1. it includes at least one node Z from S such that the arrows along the path at Z meet head to tail or tail to tail; or
2. it includes at least one node W such that the arrows along the path at W meet head to head, and neither W nor any of its descendants are in S.

Theorem

$X \perp\!\!\!\perp Y \mid S$ if and only if all paths from X to Y are blocked.

Learning parameters in Bayes nets

Recall the general form of a discrete Bayes net:

$$p(\mathbf{x}) = \prod_{v \in V} p(x_v | \mathbf{x}_{\mathsf{pa}(v)}) \qquad x_v \in \{1, 2, \ldots, k_v\}\,.$$

Assuming for now that everyone has two parents, $(x_{m(v)}, x_{f(v)})$, the conditional distributions can be parametrized by 3D arrays $\theta_1, \ldots, \theta_k$:

$$p(x_v | x_{m(v)}, x_{f(v)}) = [\theta_v]_{x_{m(v)}, x_{f(v)}, x_v}\,.$$

To ensure normalization, $\sum_{x_v} [\theta_v]_{x_{m(v)}, x_{f(v)}, x_v} = 1$ for all $x_{m(v)}, x_{f(v)}$.

Given data $\mathcal{D} = (\mathbf{x}^1, \ldots, \mathbf{x}^T)$, what is the MLE setting of $(\theta_v)_{v \in V}$?

Simpson's paradox: word of caution

You are trying to determine whether a particular treatment for a serious disease is beneficial. Given the following observations would you recommend it?

	Survived	Did not survive	Survival rate
Treatment	20	20	50%
No treatment	16	24	40%

Now what if you discovered that the breakdown by gender was this?

Males	Survived	Did not survive	Survival rate
Treatment	18	12	60%
No treatment	7	3	70%

Females	Survived	Did not survive	Survival rate
Treatment	2	8	20%
No treatment	9	21	30%

Simpson's paradox

- A graphical model can never capture all the variables that might possibly be relevant. In the first case we ignored gender. This can affect what interpretation the model suggests.
- The fact that there is an arrow from A (treatment) to B (outcome) does not imply that A causes B. In our case we had a hidden common cause, gender, of the opposite effect on B.
- To tease out causal structure we need more sophisitcated tools than just ordinary graphical models: need to introduce **interventions**.
- Observational studies are not sufficient. The gold standard in medicine is **randomized controlled trials (RCTs)**.

Learning parameters in Bayes nets

$$p(x_v|x_{m(v)}, x_{f(v)}) = [\theta_v]_{x_{m(v)}, x_{f(v)}, x_v} .$$

$$\ell(\theta|\mathcal{D}) = \prod_{t=1}^{T} \prod_{v \in V} [\theta_v]_{x^t_{m(v)}, x^t_{f(v)}, x^t_v} = \prod_{v \in V} \ell_v(\theta_v|\mathcal{D})$$

$$\ell_v(\theta_v|\mathcal{D}) = \prod_{t=1}^{T} [\theta_v]_{x^t_m, x^t_f, x^t_v}$$

or

$$\prod_a \prod_b \frac{N_{a,b}!}{N_{a,b,1}!\, N_{a,b,2}! \ldots N_{a,b,k_v}!} \; [\theta_v]_{a,b,1}^{N_{a,b,1}} \; [\theta_v]_{a,b,2}^{N_{a,b,2}} \ldots [\theta_v]_{a,b,v_k}^{N_{a,b,v_k}}$$

$$N_{a,b,c} = \left| \; \{ \; t \mid x^t_m = a, \; x^t_f = b, \; x^t_v = c \; \} \; \right|$$

Learning parameters in Bayes nets

Each

$$\ell_{v,a,b}(\theta_v|\mathcal{D}) = \frac{N_{a,b}!}{N_{a,b,1}!\,N_{a,b,2}!\ldots N_{a,b,k_v}!}\,[\theta_v]_{a,b,1}^{N_{a,b,1}}\ldots[\theta_v]_{a,b,v_k}^{N_{a,b,v_k}}$$

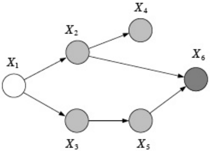
is just a multinomial like in Naive Bayes, so we know the MLE is

$$[\widehat{\theta}_v]_{a,b,c} = \frac{N_{a,b,c}}{\sum_c N_{a,b,c}}\,.$$

As before, can also use biased estimator

$$[\widehat{\theta}_v]_{a,b,c} = \frac{N_{a,b,c}+\gamma}{\sum_c (N_{a,b,c}+\gamma)}\,.$$

Inference in Bayes nets: example



The key is to factor and then apply the distributive law.

$$
\begin{aligned}
p(\mathbf{x}_1|\bar{\mathbf{x}}_6) &= p(\mathbf{x}_1, \bar{\mathbf{x}}_6)/p(\bar{\mathbf{x}}_6) \\
&= p(\mathbf{x}_1, \bar{\mathbf{x}}_6)/\sum_{\mathbf{x}_1'} p(\mathbf{x}_1', \bar{\mathbf{x}}_6)
\end{aligned}
$$

$$
\begin{aligned}
p(\mathbf{x}_1, \bar{\mathbf{x}}_6) &= \sum_{\mathbf{x}_2}\sum_{\mathbf{x}_3}\sum_{\mathbf{x}_4}\sum_{\mathbf{x}_5} p(\mathbf{x}_1)p(\mathbf{x}_2|\mathbf{x}_1)p(\mathbf{x}_3|\mathbf{x}_1)p(\mathbf{x}_4|\mathbf{x}_2)p(\mathbf{x}_5|\mathbf{x}_3)p(\bar{\mathbf{x}}_6|\mathbf{x}_2, \mathbf{x}_5) \\
&= p(\mathbf{x}_1)\sum_{\mathbf{x}_2} p(\mathbf{x}_2|\mathbf{x}_1)\sum_{\mathbf{x}_3} p(\mathbf{x}_3|\mathbf{x}_1)\sum_{\mathbf{x}_4} p(\mathbf{x}_4|\mathbf{x}_2)\sum_{\mathbf{x}_5} p(\mathbf{x}_5|\mathbf{x}_3)p(\bar{\mathbf{x}}_6|\mathbf{x}_2, \mathbf{x}_5) \\
&= p(\mathbf{x}_1)\sum_{\mathbf{x}_2} p(\mathbf{x}_2|\mathbf{x}_1)\sum_{\mathbf{x}_3} p(\mathbf{x}_3|\mathbf{x}_1)\Phi_5(\mathbf{x}_2, \mathbf{x}_3)\sum_{\mathbf{x}_4} p(\mathbf{x}_4|\mathbf{x}_2) \\
&= p(\mathbf{x}_1)\sum_{\mathbf{x}_2} p(\mathbf{x}_2|\mathbf{x}_1)\Phi_4(\mathbf{x}_2)\sum_{\mathbf{x}_3} p(\mathbf{x}_3|\mathbf{x}_1)\Phi_5(\mathbf{x}_2, \mathbf{x}_3) \\
&= p(\mathbf{x}_1)\sum_{\mathbf{x}_2} p(\mathbf{x}_2|\mathbf{x}_1)\Phi_4(\mathbf{x}_2)\Phi_3(\mathbf{x}_1, \mathbf{x}_2) \\
&= p(\mathbf{x}_1)\Phi_2(\mathbf{x}_1)
\end{aligned}
$$

Is there a general algorithm that allows us to find factorizations like this?

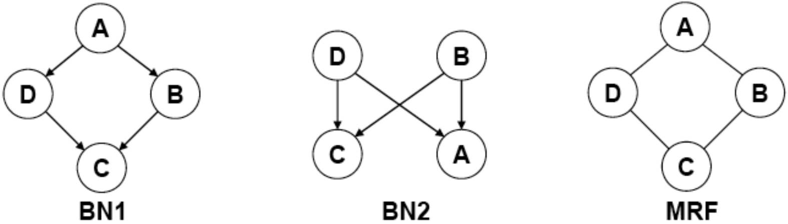
$\rightarrow$ Message passing algorithms

Undirected graphical models

Undirected graphical models

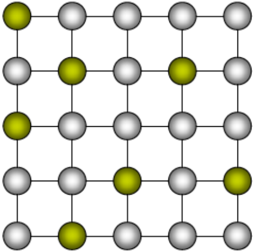
Not every type of conditional dependency structure can be represented by a Bayes net. Example:

$$X_A \perp\!\!\!\perp X_C | \{X_B, X_D\}, \qquad X_B \perp\!\!\!\perp X_D | \{X_A, X_C\}.$$

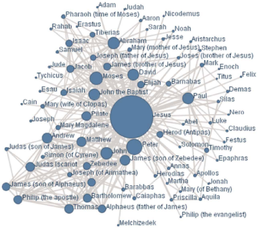


Exercise: Give an example of a structure that cannot be represented by a directed model either.

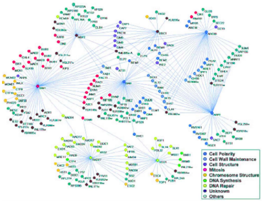
Examples of undirected models



Grid model (e.g., Ising)



Social Network



Protein interaction net

Ordinary separation

Recall the general form of the undirected models:

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{c \in \text{Cliques}(\mathcal{G})} \phi_c(\mathbf{x}_c)$$

Is X independent of Y given the set of nodes S?

Theorem

$X \perp\!\!\!\perp Y \mid S$ if and only if all paths from X to Y contain at least one node in S.

This is simpler than in the directed case.

Parameter estimation and inference

In undirected models

- Parameter estimation: Not as easy as in the directed case!
- Inference : message passing algorithms.

Bayesian vs. Frequentists

Joint and conditional probability

Joint:

$$\mathbb{P}(A, B) = \mathbb{P}(A, B)$$

Conditional:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A, B)}{\mathbb{P}(B)}$$

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A, B)}{\mathbb{P}(A)}$$

AI is all about conditional probabilities.

Mammography

Sensitivity of screening mammogram $p(+|\text{cancer}) \approx 90\%$

Specificity of screening mammogram $p(-|\text{no cancer}) \approx 91\%$

Probability that a woman age 40 has breast cancer $\approx 1\%$

If a previously unscreened 40 year old woman's mammogram is positive, what is the probability that she has breast cancer?

$$\mathbb{P}(\text{cancer}|+) = \frac{\mathbb{P}(\text{cancer}, +)}{\mathbb{P}(+)} = \frac{\mathbb{P}(+|\text{cancer})\,\mathbb{P}(\text{cancer})}{\mathbb{P}(+)} =$$

$$\frac{0.01 \times .9}{0.01 \times .9 + 0.99 \times 0.09} \approx \frac{0.009}{0.009 + 0.09} \approx \frac{0.009}{0.1} \approx 9\%$$

Message: $\mathbb{P}(A|B) \neq \mathbb{P}(B|A)$.

Bayes' rule

$$p(B|A) = \frac{p(A|B)\,p(B)}{p(A)}$$

Rev. Thomas Bayes (1701–1761)

Prosecutor's fallacy: Sally Clark

Sally Clark (1964–2007)

Two kids died with no explanation.

Sir Roy Meadow testified that chance of this happening due to SIDS is $(1/8500)^2 \approx (73 \times 10^6)^{-1}$.

Sally Clark found guilty and imprisoned.

Later verdict overturned and Meadow struck off medical register.

Fallacy: $\mathbb{P}(\text{SIDS}|2\,\text{deaths}) \neq \mathbb{P}(\text{SIDS}, 2\,\text{deaths})$

$$\mathbb{P}(\text{guilty}|+) = 1 - \mathbb{P}(\text{not guilty}|+) \neq 1 - \mathbb{P}(+|\text{not guilty})$$

Convict if ...

$$\mathbb{P}(\text{innocence}) < \texttt{reasonable doubt}$$

or

$$\mathbb{P}(\text{innocence}) < \texttt{shadow of a doubt}$$

Statistical estimation

The fundamental problem of statistics

Probability:

$$\overbrace{\theta}^{\text{parameter}} \xrightarrow{p} \overbrace{p(x)}^{\text{the model}} \xrightarrow{sampling(IID)} \overbrace{S = \{X_1, X_2, \ldots, X_m\}}^{\text{the sample}}$$

Statistics:

$$S = \{X_1, X_2, \ldots, X_m\} \xrightarrow{\text{estimation}} \widehat{\theta}$$

The fundamental problem of statistics

The problem of inferring θ from $X_1, X_2, \ldots, X_m$ is inherently ill defined:

- $p_\theta(x)$ as a function of x is a probability — OK
- $p_\theta(x)$ as a function of θ (called the **likelihood** $\ell(\theta) = p_\theta(x)$) is *not* a probability — Panic!

Two religions:

- Turn $p_\theta(x)$ into a probability by putting a distribution on θ (called the **prior**) $\rightarrow$ **Bayesian statistics**
- Just come up with a guess $\hat{\theta}$ for θ and then show that it is unlikely to get a sample that will lead to a $\hat{\theta}$ which is far off. $\rightarrow$ **Frequentist statistics**

Bayesian statistics in ML

$$\overbrace{p(\theta|X)}^{\text{posterior}} = \frac{\overbrace{p(X|\theta)}^{\text{likelihood}}\ \overbrace{p(\theta)}^{\text{prior}}}{\underbrace{p(X)}_{\text{evidence}}} = \frac{p(X|\theta)\, p(\theta)}{\int p(X|\theta)\, p(\theta)\, d\theta}$$

Example: HMM for tracking

Assumptions are steep, but at least we are honest about them.

Frequentist statistics in ML

- Come up with an algorithm to get an estimator
- Try and justify later.

Example: perceptron

Justification itself involves frequentist statistics, because it requires *estimating* the probability of error! $\rightarrow$ Error analysis of Bayesian methods also requires frequentist statistics.

Frequentist vs. Bayesian estimators

In the more classical setting of parametric models, how do we get an actual estimator for θ?

Frequentist:

- Use the maximum likelihood estimator $\widehat{\theta}_{\text{MLE}} = \arg\max \ell(\theta)$
- Just one of many options

Bayesian:

- A true Bayesian always reports the full posterior $p(\theta|X)$.
- When pressed, might give the maximum a posteriori (MAP) estimator $\hat{\theta}_{\text{MAP}} = \arg\max p(\theta|X)$
- or the posterior mean $\hat{\theta} = \int \theta \, p(\theta|X) \, d\theta$.

Naive Bayes

A generative model for documents

Let x_i be the number of times that word i occurs in a document $\mathcal{D}$.

Generative model for docs from author A: $p_A(\mathbf{x})$
Generative model for docs from author B: $p_B(\mathbf{x})$

Given a new document with vector $\mathbf{x}'$ attribute it to A iff

$$p_A(\mathbf{x}') > p_B(\mathbf{x}')$$

What form should p take?

The multinomial model

Mutinomial (Naive Bayes) model for word counts:

$$p(\mathbf{x}) = \frac{n!}{x_1!\, x_2! \ldots x_k!}\, \theta_1^{x_1}\, \theta_2^{x_2} \ldots \theta_k^{x_k} \qquad \sum_{i=1}^{k} \theta_i = 1.$$

How do we find the parameters $\theta_1, \theta_2, \ldots, \theta_k$?

Naive Bayes — the Frequentist way

$$p(\mathbf{x}) = \frac{n!}{x_1!\, x_2! \ldots x_k!}\, \theta_1^{x_1}\, \theta_2^{x_2} \ldots \theta_k^{x_k} \qquad \sum_{i=1}^{k} \theta_i = 1.$$

MLE: maximize

$$\log \ell(\theta_1, \theta_2, \ldots, \theta_k) = \text{constant} + \sum_{i=1}^{k} x_i \log \theta_i \quad \text{s.t.} \quad \sum_{i=1}^{k} \theta_i = 1$$

Introduce the λ **Lagrange multiplier**:

$$\frac{\partial}{\partial \theta_i} \left[\log \ell(\theta_1, \ldots, \theta_k) + \lambda \sum_{i=1}^{k} \theta_i \right] = 0$$

$$\frac{x_i}{\widehat{\theta}_i} = \lambda \quad \longrightarrow \quad \widehat{\theta}_i = \frac{x_i}{\sum_{i=1}^{k} x_i}$$

Naive Bayes — the Frequentist way

The MLE $\theta_i = x_i / \sum_{i=1}^{k} x_i$ makes perfect sense but gives $p(\mathbf{x}') = 0$ whenever $\mathcal{D}'$ contains a word not seen in training!

Idea: **bias** the estimator:

$$\widehat{\theta}_i = \frac{x_i + \gamma}{k\gamma + \sum_{i=1}^{k} x_i}$$

where γ is a "pseudocount".

Naive Bayes — the Bayesian way

1. Take the prior $p(\theta_1, \theta_2, \dots, \theta_k)$ to be

$$\text{Dirichlet}(\alpha_1, \alpha_2, \dots, \alpha_k) = \frac{\Gamma(\alpha_1 + \dots + \alpha_k)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_k)} \theta_1^{\alpha_1 - 1} \dots \theta_k^{\alpha_k - 1}$$

   where Γ is the Gamma function obeying $\Gamma(n) = (n-1)!$ for any $n \in \mathbb{N}$.

2. Apply Bayes' rule

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\int p(\mathbf{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\theta}$$

3. The posterior becomes

$$p(\theta_1, \theta_2, \dots, \theta_k | \boldsymbol{X}) = \text{Dirichlet}(\alpha_1 + x_1, \dots, \alpha_k + x_k).$$

Naive Bayes — the Bayesian way

The maximum a posterior (MAP) is given by

$$\theta_i = \frac{x_i + \alpha_i - 1}{\sum_{i=1}^{k} x_i + \alpha_i - 1}$$

Equivalent to frequentist estimator with pseudocounts of $\alpha_i - 1$!!!

Learning the parameters of Bayes nets

Parameter Learning

Recall that the joint distribution of Bayes net is of the form

$$p(\mathbf{x}) = \prod_{v \in V} p_v(x_v | \mathbf{x}_{\mathsf{pa}(v)}).$$

Up to now, we have assumed that each p_v is fully specified, and asked questions about the distribution of certain variables given others.

Now the question is:

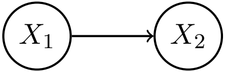
- Assuming that each $p_v(x_v | \mathbf{x}_{\mathsf{pa}(v)})$ is parametrized by some set of parameters Θ_v , given a training set $\{\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^m\}$ of all the variables together, how should set the Θ_v parameters?

This type of task is central to using graphical models in practice.

A simple model

The simplest case is a model with just two variables, $\mathbf{x} = (x_1, x_2)$:

$$p(\mathbf{x}) = p(x_2|x_1)\, p(x_1)$$



Assume that:

- x_1 can take on k_1 different values $\{1, 2, \dots, k_1\}$,
- x_2 can take on k_2 different values $\{1, 2, \dots, k_2\}$.

In this case $p(x_2|x_1)$ is described by the matrix $\Theta \in [0,1]^{k_1 \times k_2}$ of parameters

$$\theta_{i,j} = p(X_2 = j | X_1 = i).$$

How do we learn this matrix from the training data $\{(x_1^u, x_2^u)\}_{u=1}^m$?

The Maximum Likelihood Principle

Assume that we are given

- a parametric family of distributions $p_\Theta(x)$ parametrized some the (set of) parameters Θ,
- a sample $\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^m\}$ from p_Θ.

The **likelihood** of $\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^m\}$ under p_Θ is

$$\ell(\Theta) = \prod_{u=1}^{m} p_\Theta(\mathbf{x}^u).$$

The **Maximum Likelihood Estimator (MLE)** $\widehat{\Theta}$ of Θ is then the setting of the parameters that maximizes this expression.

Finding (estimating) parameters is a general strategy in statistics, not just for Bayes nets. Often it is slightly more convenient to maximize the **log-likelihood**

$$\log \ell(\Theta) = \sum_{u=1}^{m} \log p_\Theta(\mathbf{x}^u).$$

Maximum Likelihood

In our case, recalling that $p(X_2 = j | X_1 = i) = \theta_{i,j}$,

$$\ell(\Theta) = \prod_{u=1}^{m} p(x_2^u | x_1^u) = \prod_{u=1}^{m} \theta_{x_1^u, x_2^u} = \prod_{i=1}^{k_1} \prod_{j=1}^{k_2} \theta_{i,j}^{N_i(j)},$$

where $N_i(j)$ is the number of samples where $X_1 = i$ and $X_2 = j$.

- The likelihood only depends on how many times we have observed each combination of (x_1, x_2) together. This makes sense.
- For each possible i , the conditional $p(X_2 = j | X_1 = i) = \theta_{i,j}$ is a prob. distr. as a function of j . $\rightarrow$ We can learn each row of Θ separately, with the constraint that $\sum_{j=1}^{k_2} \theta_{i,j} = 1$.

Maximum Likelihood

To learn $(\theta_{i,1}, \theta_{i,2}, \ldots, \theta_{i,k_2})$ only need the likelihood over those training examples in which $x_1 = i$:

$$\ell_i(\theta_{i,1}, \ldots, \theta_{i,k_2}) = \prod_{j=1}^{k_2} \theta_{i,j}^{N_i(j)}.$$

In logarithmic form, to get the MLE $(\widehat{\theta}_{i,1}, \widehat{\theta}_{i,2}, \ldots, \widehat{\theta}_{i,k_2})$, solve

$$\text{maximize} \sum_{j=1}^{k_2} N_i(j) \log \theta_{i,j} \qquad \text{subject to} \qquad \sum_{j=1}^{k_2} \theta_{i,j} = 1.$$

This is a constrained optimization problem. $\rightarrow$ Use Lagrange multipliers.

Solving the optimization problem

The Lagrangian is

$$\mathcal{L}(\theta_{i,1}, \ldots, \theta_{i,k_2}; \lambda) = \sum_{j=1}^{k_2} N_i(j) \log \theta_{i,j} - \lambda \sum_{j=1}^{k_2} \theta_{i,j}.$$

At the optimum:

$$\frac{\partial}{\partial \theta_{i,j}} \mathcal{L}(\ldots) = 0 \quad \Longrightarrow \quad N_i(j) \frac{1}{\theta_{i,j}} - \lambda = 0 \quad \Longrightarrow \quad \theta_{i,j} = \frac{N_i(j)}{\lambda}.$$

Impose the constraint:

$$\sum_{j=1}^{k_2} \frac{N_i(j)}{\lambda} = 1 \quad \Longrightarrow \quad \lambda = \sum_{j=1}^{k_2} N_i(j) \quad \Longrightarrow \quad \boxed{\widehat{\theta}_{i,j} = \frac{N_i(j)}{\sum_{j'=1}^{k_2} N_i(j')}}$$

This solution also makes sense intuitively. But what if $N_i(j) = 0$?

Zero counts

One problem with the MLE

$$\widehat{\theta}_{i,j} = \frac{N_i(j)}{\sum_{j'=1}^{k_2} N_i(j')}$$

is that if one (i, j) pair never occurs in the training data, then the corresponding $\widehat{\theta}_{i,j}$ will be zero. $\rightarrow$ The learned model will not be able to deal with any example in which $X_1 = i$ and $X_2 = j$.

Standard solution: add a small "pseudocount" to each $N_i(j)$, e.g., $\gamma = 0.1$:

$$\widehat{\theta}_{i,j} = \frac{N_i(j) + \gamma}{\sum_{j'=1}^{k_2}(N_i(j') + \gamma)} .$$

This is a form of **regularization.** We will see other interpretations later.

Multiple parents

What if the model is

$$p(\mathbf{x}) = p(x_{r+1}|x_1, x_2, \ldots, x_r)\,?$$

Now Θ is a $k_1 \times \ldots \times k_r \times k_{r+1}$ array (tensor) with

$$\theta_{i_1,\ldots,i_r,j} = p(\,X_{r+1} = j \mid X_1 = i_1, \ldots, X_r = i_r\,).$$

However, fixing any $(i_1, \ldots, i_p)$ we can solve for the corresponding $\widehat{\theta}_{\ldots,j}$'s just as before and get (using pseudocounts)

$$\widehat{\theta}_{i_1,\ldots,i_r,j} = \frac{N_{i_1,\ldots,i_r}(j) + \gamma}{\sum_{j'=1}^{k_{r+1}} (N_{i_1,\ldots,i_r}(j') + \gamma)}\,.$$

MLE for the whole Bayes Net

Recall that the joint distribution of the whole Bayes net is

$$p(\mathbf{x}) = \prod_{v \in V} p_v(x_v | \mathbf{x}_{\mathsf{pa}(v)}),$$

so we need to learn a separate Θ_v array for each of these factors. Assuming that each of the variables is discrete, this is not so hard. For each v, assuming that the parents of v are $\{p_1, \ldots, p_r\}$:

1. Form the corresponding likelihood

$$\ell(\Theta_v) = \prod_{u=1}^{m} p(x_v^u | x_{p_1}^u, \ldots, x_{p_r}^u).$$

2. Use the same steps as before to find the corresponding MLE solution

$$[\widehat{\theta}_v]_{i_1,\ldots,i_r,j} = \frac{N_{i_1,\ldots,i_r}(j) + \gamma}{\sum_{j'=1}^{k_{r+1}} (N_{i_1,\ldots,i_r}(j') + \gamma)} .$$

Expectation maximization

What about when some of the x_i's are not observed? Use the EM strategy and iterate until convergence:

1. **E-step:** Compute the *expected* log-likelihood (w.r.t. the hidden variables) under $\widehat{\Theta}_{\text{old}}$
$$\overline{\mathcal{L}}_{\widehat{\Theta}_{\text{old}}}(\Theta).$$
2. **M-step:** Maximize this to get the new estimate for $\widehat{\Theta}$:
$$\widehat{\Theta} = \arg\max_{\Theta} \overline{\mathcal{L}}_{\widehat{\Theta}_{\text{old}}}(\Theta).$$

Just like in probabilistic k–means. Whether or not this is viable for a complicated model is not obvious.

FURTHER READING

- David Barber: **Bayesian Reasoning and Machine Learning** (online)
- Daphne Koller and Nir Friedman: **Probabilistic Graphical Models**
- Tutorial by Sam Roweis: `http://videolectures.net/mlss06tw_roweis_mlpgm/`
- Coursera course “Probabilistic Graphical Models” by Daphne Koller