CS 1571 Introduction to AI
Lecture 25

Learning probability distributions

Milos Hauskrecht
milos@cs.pitt.edu
5329 Sennott Square

Unsupervised learning

- **Data:** $D = \{D_1, D_2, .., D_n\}$
  $D_i = \mathbf{x}_i$ a vector of attribute values
  - e.g. the description of a patient
  - no specific target attribute we want to predict (no output y)
- **Objective:**
  - learn (describe) relations between attributes, examples

Types of problems:
- **Clustering**
  Group together "similar" examples
- **Density estimation**
  - Model probabilistically the population of examples

Density estimation

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

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

Attributes:

- modeled by random variables $\mathbf{X} = \{X_1, X_2, \ldots, X_d\}$ with:
  - **Continuous values**
  - **Discrete values**

  E.g. ***blood pressure*** with numerical values

  or ***chest pain*** with discrete values

  [no-pain, mild, moderate, strong]

Underlying true probability distribution:

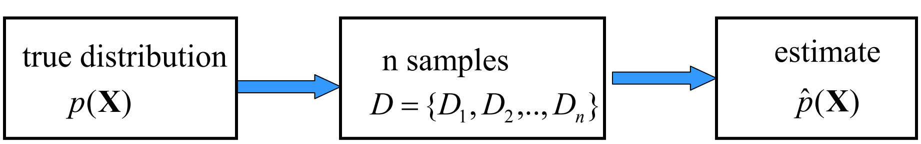
$p(\mathbf{X})$

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 $p(\mathbf{X})$)**

Learning via parameter estimation

In this lecture we consider **parametric density estimation**

Basic settings:

- A set of random variables $\mathbf{X} = \{X_1, X_2, \dots, X_d\}$
- **A model of the distribution** over variables in $\boldsymbol{X}$ with parameters Θ
- **Data** $D = \{D_1, D_2, .., D_n\}$

Objective: find parameters $\hat{\Theta}$ that fit the data the best, or in other words reduce the misfit between the data and the model

- What is the best set of parameters?
  - There are various criteria one can apply here.

Parameter estimation. Basic criteria.

- **Maximum likelihood (ML) criterion**

$$\arg\max_{\Theta} p(D \mid \Theta, \xi) \quad \leftarrow \text{Likelihood of data}$$

ξ - represents prior (background) knowledge

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

$$\arg\max_{\Theta} p(\Theta \mid D, \xi) \quad \leftarrow \text{Posterior probability}$$

MAP selects the mode of the posterior

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

Parameter estimation. Coin example.

Coin example: we have a coin that can be biased

Outcomes: two possible values -- head or tail

Data: D a sequence of outcomes x_i such that

- **head** $x_i = 1$
- **tail** $x_i = 0$

Model: probability of a head θ

probability of a tail $(1-\theta)$

Objective:

We would like to estimate the probability of a **head** $\hat{\theta}$ from data

Parameter estimation. Example.

- **Assume** the unknown and possibly biased coin
- Probability of the head is θ
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10

What would be your estimate of the probability of a head ?

$$\tilde{\theta} = ?$$

Parameter estimation. Example

- **Assume** the unknown and possibly biased coin
- Probability of the head is θ
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10

What would be your choice of the probability of a head ?

Solution: use frequencies of occurrences to do the estimate

$$\tilde{\theta} = \frac{15}{25} = 0.6$$

This is **the maximum likelihood estimate** of the parameter θ

Probability of an outcome

Data: D a sequence of outcomes x_i such that

- **head** $x_i = 1$
- **tail** $x_i = 0$

Model: probability of a head θ

probability of a tail $(1-\theta)$

Assume: we know the probability θ

Probability of an outcome of a coin flip x_i

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

- Combines the probability of a head and a tail
- So that x_i is going to pick its correct probability
- Gives θ for $x_i = 1$
- Gives $(1-\theta)$ for $x_i = 0$

Probability of a sequence of outcomes.

Data: D a sequence of outcomes x_i such that

- **head** $x_i = 1$
- **tail** $x_i = 0$

Model: probability of a head θ
probability of a tail $(1-\theta)$

Assume: a sequence of independent coin flips

D = H H T H T H (encoded as D= 110101)

What is the probability of observing the data sequence **D:**

$$P(D \mid \theta) = ?$$

Probability of a sequence of outcomes.

Data: D a sequence of outcomes x_i such that

- **head** $x_i = 1$
- **tail** $x_i = 0$

Model: probability of a head θ
probability of a tail $(1-\theta)$

Assume: a sequence of coin flips D = H H T H T H encoded as D= 110101

What is the probability of observing a data sequence **D:**

$$P(D \mid \theta) = \theta\theta(1-\theta)\theta(1-\theta)\theta$$

Probability of a sequence of outcomes.

Data: D a sequence of outcomes x_i such that

- **head** $x_i = 1$
- **tail** $x_i = 0$

Model: probability of a head θ
probability of a tail $(1-\theta)$

Assume: **a sequence of coin flips D = H H T H T H**
encoded as D= 110101

What is the probability of observing a data sequence **D:**

$$P(D \mid \theta) = \theta\theta(1-\theta)\theta(1-\theta)\theta$$

likelihood of the data

Probability of a sequence of outcomes.

Data: D a sequence of outcomes x_i such that

- **head** $x_i = 1$
- **tail** $x_i = 0$

Model: probability of a head θ
probability of a tail $(1-\theta)$

Assume: **a sequence of coin flips D = H H T H T H**
encoded as D= 110101

What is the probability of observing a data sequence **D:**

$$P(D \mid \theta) = \theta\theta(1-\theta)\theta(1-\theta)\theta$$

$$P(D \mid \theta) = \prod_{i=1}^{6} \theta^{x_i}(1-\theta)^{(1-x_i)}$$

Can be rewritten using the Bernoulli distribution:

The goodness of fit to the data.

Learning: we do not know the value of the parameter θ

Our learning goal:

- Find the parameter θ that fits the data D the best?

One solution to the "best": Maximize the likelihood

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

Intuition:

- more likely are the data given the model, the better is the fit

Note: Instead of an error function that measures how bad the data fit the model we have a measure that tells us how well the data fit :

$$Error\ (D, \theta) = -P(D \mid \theta)$$

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 (the same as maximizing likelihood)

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

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

N_1 - number of heads seen N_2 - number of tails 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 likelihood estimate. Example

- **Assume** the unknown and possibly biased coin
- Probability of the head is θ
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10

What is the ML estimate of the probability of a head and a tail?

Maximum likelihood estimate. Example

- Assume the unknown and possibly biased coin
- Probability of the head is θ
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10

What is the ML estimate of the probability of head and tail ?

Head: $$\theta_{ML} = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2} = \frac{15}{25} = 0.6$$

Tail: $$(1 - \theta_{ML}) = \frac{N_2}{N} = \frac{N_2}{N_1 + N_2} = \frac{10}{25} = 0.4$$

Maximum a posteriori estimate

Maximum a posteriori estimate

- Selects the mode of the **posterior distribution**

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

Likelihood of data → $P(D \mid \theta, \xi)$; **prior** → $p(\theta \mid \xi)$

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

Normalizing factor → $P(D \mid \xi)$

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

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

How to choose the prior probability?

Prior distribution

Choice of prior: Beta distribution

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

$\Gamma(x)$ - A Gamma function

For integer values of x $\Gamma(x) = x!$

Why to use Beta distribution?

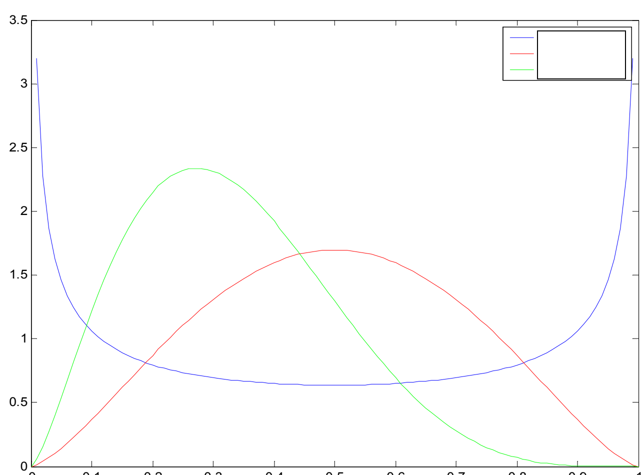
Beta distribution "**fits**" Bernoulli trials - **conjugate choices**

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

Posterior distribution is again a Beta distribution

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

Beta distribution



Maximum a posterior probability

Maximum a posteriori estimate

– Selects the mode of the **posterior distribution**

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

$$= \frac{\Gamma(\alpha_1 + \alpha_2 + N_1 + N_2)}{\Gamma(\alpha_1 + N_1)\Gamma(\alpha_2 + N_2)} \theta^{N_1+\alpha_1-1}(1-\theta)^{N_2+\alpha_2-1}$$

Notice that parameters of the prior act like counts of heads and tails (sometimes they are also referred to as **prior counts**)

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

MAP estimate example

- Assume the unknown and possibly biased coin
- Probability of the head is θ
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10
- Assume $p(\theta \mid \xi) = Beta(\theta \mid 5,5)$

What is the MAP estimate?

MAP estimate example

- Assume the unknown and possibly biased coin
- Probability of the head is θ
- **Data:**
  H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10
- Assume $p(\theta \mid \xi) = Beta(\theta \mid 5,5)$

What is the MAP estimate ?

$$\theta_{MAP} = \frac{N_1 + \alpha_1 - 1}{N - 2} = \frac{N_1 + \alpha_1 - 1}{N_1 + N_2 + \alpha_1 + \alpha_2 - 2} = \frac{19}{33}$$

MAP estimate example

- Note that the prior and data fit (data likelihood) are combined
- **The MAP can be biased with large prior counts**
- **It is hard to overturn it with a smaller sample size**
- **Data:**
  H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10
- Assume

$$p(\theta \mid \xi) = Beta(\theta \mid 5,5) \qquad \theta_{MAP} = \frac{19}{33}$$

$$p(\theta \mid \xi) = Beta(\theta \mid 5,20) \qquad \theta_{MAP} = \frac{19}{48}$$

Multinomial distribution

Example: **Multi-way coin toss, roll of dice**

- **Data:** a set of N outcomes (multi-set)

  N_i - a number of times an outcome i has been seen

Model parameters: $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots \theta_k)$ **s.t.** $\sum_{i=1}^{k} \theta_i = 1$

θ_i - probability of an outcome i

Probability of data (likelihood)

$$P(N_1, N_2, \ldots N_k \mid \boldsymbol{\theta}, \xi) = \frac{N!}{N_1! N_2! \ldots N_k!} \theta_1^{N_1} \theta_2^{N_2} \ldots \theta_k^{N_k}$$

Multinomial distribution

ML estimate:

$$\theta_{i,ML} = \frac{N_i}{N}$$

MAP estimate

Choice of prior: Dirichlet distribution

$$Dir(\boldsymbol{\theta} \mid \alpha_1, .., \alpha_k) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \ldots \theta_k^{\alpha_k - 1}$$

Dirichlet is the conjugate choice for multinomial

$$P(D \mid \boldsymbol{\theta}, \xi) = P(N_1, N_2, \ldots N_k \mid \boldsymbol{\theta}, \xi) = \frac{N!}{N_1! N_2! \ldots N_k!} \theta_1^{N_1} \theta_2^{N_2} \ldots \theta_k^{N_k}$$

Posterior distribution

$$p(\boldsymbol{\theta} \mid D, \xi) = \frac{P(D \mid \boldsymbol{\theta}, \xi) Dir(\boldsymbol{\theta} \mid \alpha_1, \alpha_2, .. \alpha_k)}{P(D \mid \xi)} = Dir(\boldsymbol{\theta} \mid \alpha_1 + N_1, .., \alpha_k + N_k)$$

MAP estimate:

$$\theta_{i,MAP} = \frac{\alpha_i + N_i - 1}{\sum_{i=1,..k} (\alpha_i + N_i) - k}$$

Learning complex distributions

- **The problem of learning complex distributions**
  - can be sometimes reduced to the problem of learning a number of simpler distributions

- Such a decomposition occurs for example in **Bayesian networks**
  - Builds upon independences encoded in the network

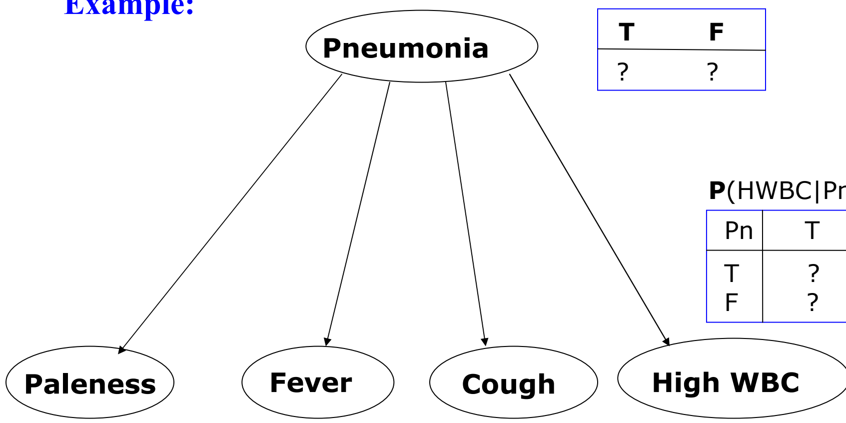
- **Why learning of BBNs?**
  - Large databases are available
    - uncover important probabilistic dependencies from data and use them in inference tasks

Learning of BBN parameters

Learning. Two steps:
  - Learning of the network structure
  - Learning of parameters of conditional probabilities
- **Variables**:
  - Observable – values present in every data sample
  - Hidden – values are never in the sample
  - Missing values – values sometimes present, sometimes not
- **Here:**
  - learning parameters for the fixed graph structure
  - All variables are observed in the dataset

Learning of BBN parameters. Example.

Example:



P(Pneumonia)

T	F
?	?

P(HWBC|Pneum)

Pn	T	F
T	?	?
F	?	?

P(Palen|Pneum) ?

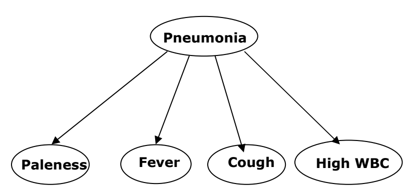
P(Fever|Pneum) ?

P(Cough|Pneum) ?

Learning of BBN parameters. Example.

Data D (different patient cases):

Pal	Fev	Cou	HWB	Pneu
T	T	T	T	F
T	F	F	F	F
F	F	T	T	T
F	F	T	F	T
F	T	T	T	T
T	F	T	F	F
F	F	F	F	F
T	T	F	F	F
T	T	T	T	T
F	T	F	T	T
T	F	F	T	F
F	T	F	F	F



Estimates of parameters of BBN

- Much like multiple **coin tosses or rolls of a dice**
- A "smaller" learning problem corresponds to the learning of exactly one conditional distribution
- **Example:**

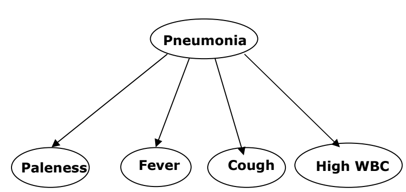
$$\mathbf{P}(Fever \mid Pneumonia = T)$$

- **Problem:** How to pick the data to learn?

Learning of BBN parameters. Example.

Data D (different patient cases):

Pal	Fev	Cou	HWB	Pneu
T	T	T	T	F
T	F	F	F	F
F	F	T	T	T
F	F	T	F	T
F	T	T	T	T
T	F	T	F	F
F	F	F	F	F
T	T	F	F	F
T	T	T	T	T
F	T	F	T	T
T	F	F	T	F
F	T	F	F	F



How to estimate:

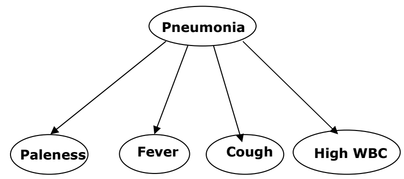
$$\mathbf{P}(Fever \mid Pneumonia = T) = ?$$

Learning of BBN parameters. Example.

Learn: $\mathbf{P}(Fever \mid Pneumonia = T)$

Step 1: Select data points with Pneumonia=T

Pal	Fev	Cou	HWB	Pneu
T	T	T	T	F
T	F	F	F	F
F	F	T	T	T
F	F	T	F	T
F	T	T	T	T
T	F	T	F	F
F	F	F	F	F
T	T	F	F	F
T	T	T	T	T
F	T	F	T	T
T	F	F	T	F
F	T	F	F	F

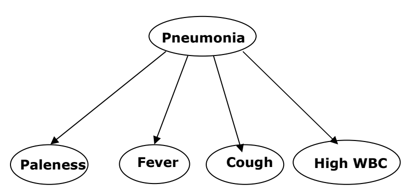


Learning of BBN parameters. Example.

Learn: $\mathbf{P}(Fever \mid Pneumonia = T)$

Step 1: Ignore the rest

Pal	Fev	Cou	HWB	Pneu
F	F	T	T	T
F	F	T	F	T
F	T	T	T	T
T	T	T	T	T
F	T	F	T	T

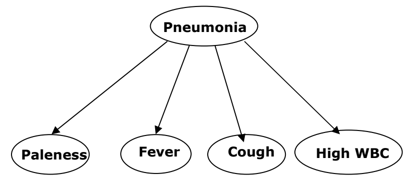


Learning of BBN parameters. Example.

Learn: $\mathbf{P}(Fever \mid Pneumonia = T)$

Step 2: Select values of the random variable defining the distribution of Fever

Pal	Fev	Cou	HWB	Pneu
F	F	T	T	T
F	F	T	F	T
F	T	T	T	T
T	T	T	T	T
F	T	F	T	T

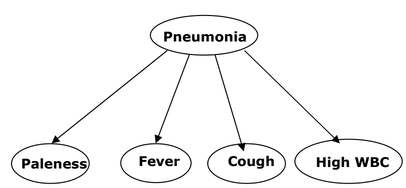


Learning of BBN parameters. Example.

Learn: $\mathbf{P}(Fever \mid Pneumonia = T)$

Step 2: Ignore the rest

Fev
F
F
T
T
T

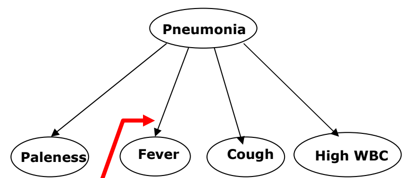


Learning of BBN parameters. Example.

Learn: $\mathbf{P}(Fever \mid Pneumonia = T)$

Step 3: Learning the ML estimate

Fev
F
F
T
T
T



$\mathbf{P}(Fever \mid Pneumonia = T)$

T	F
0.6	0.4

Learning of BBN parameters. Example.

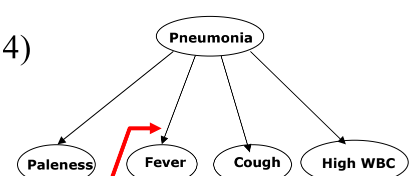
Learn: $\mathbf{P}(Fever \mid Pneumonia = T)$

Step 3: Learning the MAP estimate

Assume the prior

$$\theta_{Fever|Pneumonia=T} \sim Beta(3,4)$$

Fev
F
F
T
T
T



$\mathbf{P}(Fever \mid Pneumonia = T)$

T	F
0.5	0.5