

CS 2710 Foundations of AI
Lecture 25

Learning probability distributions

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

CS 2710 Foundations of AI

Density estimation

Data: $D = \{D_1, D_2, \dots, D_n\}$
 $D_i = \mathbf{x}_i$ a vector of attribute values

Attributes:

- modeled by random variables $\mathbf{X} = \{X_1, X_2, \dots, X_d\}$ 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:

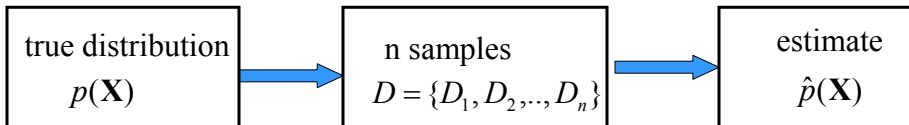
$$p(\mathbf{X})$$

CS 2710 Foundations of AI

Density estimation

Data: $D = \{D_1, D_2, \dots, D_n\}$
 $D_i = \mathbf{x}_i$ 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 \mathbf{X} with parameters Θ
- **Data** $D = \{D_1, D_2, \dots, D_n\}$

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**

$\arg \max_{\Theta} p(D | \Theta, \xi)$ ← Likelihood of data

ξ - represents prior (background) knowledge

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

$\arg \max_{\Theta} p(\Theta | D, \xi)$ ← Posterior probability

MAP selects the mode of the posterior

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

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:

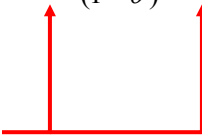
We would like to estimate the probability of a **head** $\hat{\theta}$
from data

Maximum a posterior probability

Maximum a posteriori estimate

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

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

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


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

- 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

$$p(\theta | \xi) = \text{Beta}(\theta | 5, 5) \quad \theta_{MAP} = \frac{19}{33}$$

$$p(\theta | \xi) = \text{Beta}(\theta | 5, 20) \quad \theta_{MAP} = \frac{19}{48}$$

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: $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)$ s.t. $\sum_{i=1}^k \theta_i = 1$
 θ_i - probability of an outcome i

Probability of data (likelihood)

$$P(N_1, N_2, \dots, N_k | \boldsymbol{\theta}, \xi) = \frac{N!}{N_1! N_2! \dots N_k!} \theta_1^{N_1} \theta_2^{N_2} \dots \theta_k^{N_k} \quad \text{Multinomial distribution}$$

ML estimate:

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

MAP estimate

Choice of prior: Dirichlet distribution

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

Dirichlet is the conjugate choice for multinomial

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

Posterior distribution

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

MAP estimate:

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

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

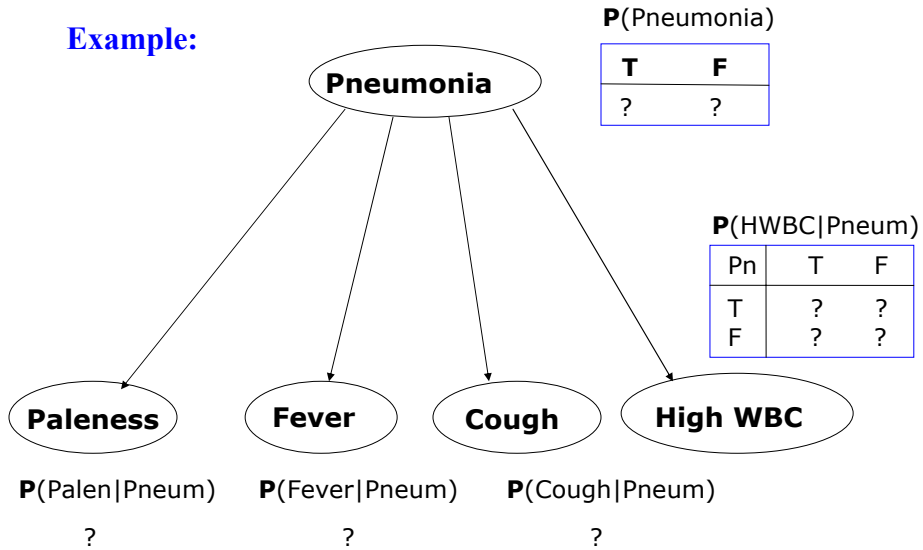
Learning of BBN parameters

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:



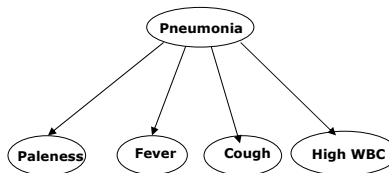
CS 2710 Foundations of AI

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



CS 2710 Foundations of AI

Estimates of parameters of BBN

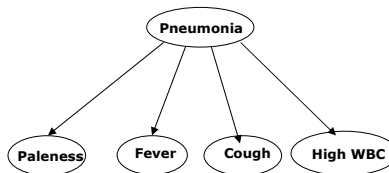
- Much like multiple **coin tosses or rolls of a dice**
- A “smaller” learning problem corresponds to the learning of exactly one conditional distribution
- **Example:** $P(\text{Fever} \mid \text{Pneumonia} = T)$
- **Problem:** How to pick the data to learn?

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:

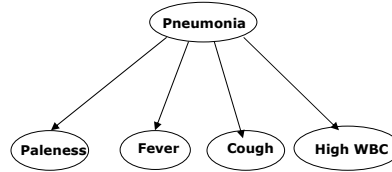
$$P(\text{Fever} \mid \text{Pneumonia} = T) = ?$$

Learning of BBN parameters. Example.

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

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

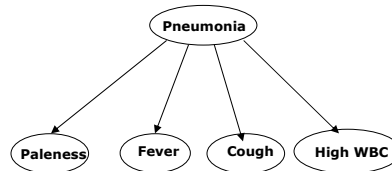


Learning of BBN parameters. Example.

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

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



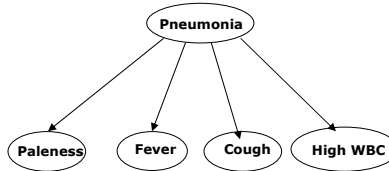
Learning of BBN parameters. Example.

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

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



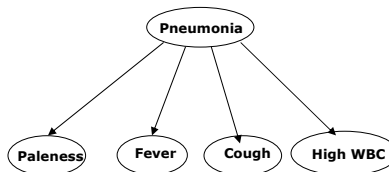
Learning of BBN parameters. Example.

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

Step 2: Ignore the rest

Fev

F
F
T
T
T



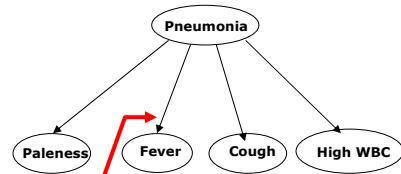
Learning of BBN parameters. Example.

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

Step 3: Learning the ML estimate

Fev

F
F
T
T
T



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

T	F
0.6	0.4

Learning of BBN parameters. Example.

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

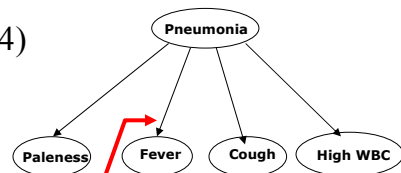
Step 3: Learning the MAP estimate

Assume the prior

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

Fev

F
F
T
T
T



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

T	F
0.5	0.5

Estimates of parameters of BBN: summary

Much like multiple **coin toss or roll of a dice** problems.

- 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?

Answer:

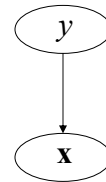
1. Select data points with Pneumonia=T
(ignore the rest)
2. Focus on (select) only values of the random variable defining the distribution (Fever)
3. Learn the parameters of the conditional the same way as we learned the parameters of the biased coin or dice

Using an unsupervised methods to do classification Generative approach to classification

Idea:

1. **Represent and learn the distribution** $p(\mathbf{x}, y)$
2. **Use it to define compute**

$$p(y = 0 \mid \mathbf{x}) \text{ and } p(y = 1 \mid \mathbf{x})$$



Typical model $p(\mathbf{x}, y) = p(\mathbf{x} \mid y)p(y)$

- $p(\mathbf{x} \mid y) =$ **Class-conditional distributions**

binary classification: two class-conditional distributions

$$p(\mathbf{x} \mid y = 0) \quad p(\mathbf{x} \mid y = 1)$$

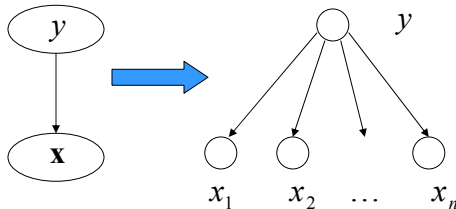
- $p(y) =$ **Priors on classes** - probability of class y

binary classification: Bernoulli distribution

$$p(y = 0) + p(y = 1) = 1$$

Naïve Bayes classifier

- a generative classifier model with an additional simplifying assumption:
 - All input attributes are conditionally independent of each other given the class. So we have:



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

$$p(\mathbf{x} | y) = \prod_{i=1}^n p(x_i | y)$$

Learning parameters of the NB model

Much simpler density estimation problems

- We need to learn:
$$p(\mathbf{x} | y = 0) \quad \text{and} \quad p(\mathbf{x} | y = 1) \quad \text{and} \quad p(y)$$
- Because of the assumption of the conditional independence we need to learn:
for every variable i : $p(x_i | y = 0)$ and $p(x_i | y = 1)$

Advantages:

- **Easy to learn if the number of input attributes is large**
- **Gives us a flexibility to represent input attributes different of different forms !!!**
 - E.g. one attribute can be modeled using the Bernoulli, the other as Gaussian density, or as a Poisson distribution

Final exam covers

- Chapters 1-9. All.
- Chapter 10: Exclude Sections 10.7. -10.8
- Chapter 11: Exclude Sections 11.4.-11.5.
- Chapter 12: Only Section 12.1-12.4 (first part)
- Chapter 13. All.
- Chapter 14. Exclude MCMC.
- Chapter 15. Exclude.
- Chapter 16. Exclude 16.5 (Decision networks)
- Chapter 17. Exclude.
- Chapter 18. Exclude 18.4 and 18.5
- Chapter 19. Exclude.
- Chapter 20. Exclude 20.3 and 20.4 and 20.6.