



2D1431 Machine Learning

# BAYESIAN LEARNING

Danica Kragic



November 13, 2006

2D1431 Machine Learning

Danica Kragic



## The Lady and the Tiger

A young prince and princess had fallen in love, but the girl's father, a bitter old King, opposed the marriage. So the King contrived to lure the Prince into a trap. In front of his entire court, he challenged the Prince to prove his love in a highly unusual and dangerous game.

"The Princess," said the King, "is behind one of these three doors I have placed in front of you. Behind the other two are hungry tigers who will most certainly eat you. If you prove your love by picking the correct door, you may marry my daughter!"

"And just to demonstrate that I'm not a bitter old man," said the King "I will help you. Once you make your choice, I will show you a tiger behind one of the other doors. And then," intoned the King, "you may pick again!". The King smiled, convinced that the Prince would not be man enough to take the challenge.

Now the Prince knew that if he walked away he would never to see his love again. So he swallowed hard, uttered a short prayer for luck, and then picked a door at random. "I choose this door," said the Prince.

"Wait!" commanded the King. "I am as good as my word. Now I will show you a tiger. Guards! Three of the Kings guards cautiously walked over to one of the other doors, opened it. A huge hungry tiger had been crouching behind it!

"Now," said the King, "Make your choice!" And, glancing to his court, he added, "Unless of course you wish to give up now and walk away.."

What should Prince do?

2D1431 Machine Learning

Danica Kragic



## Introduction

- Bayesian Decision Theory came long before Version Spaces, Decision Tree Learning and Neural Networks. It was studied in the field of Statistical Theory and more specifically, in the field of Pattern Recognition.
- Bayesian Decision Theory is at the basis of important learning schemes such as the Naïve Bayes Classifier, Learning Bayesian Belief Networks and the EM Algorithm.
- Bayesian Decision Theory is also useful as it provides a framework within which many non-Bayesian classifiers can be studied (See [Mitchell, Sections 6.3, 4,5,6]).

2D1431 Machine Learning

Danica Kragic



## Discrete Random Variables

- $A$  is a Boolean-valued random variable if it denotes an event and there is some degree of uncertainty as to whether  $A$  occurs
- Examples:
  - $A$  = The SK1001 pilot is a male
  - $A$  = Tomorrow will be a sunny day
  - $A$  = You will master today's lecture

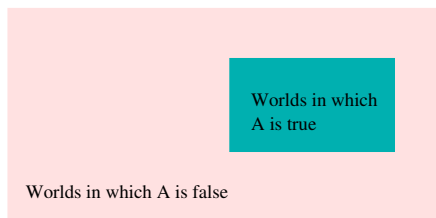
2D1431 Machine Learning

Danica Kragic



## Probabilities

- $P(A)$  - "fraction of all possible worlds in which  $A$  is true"
- $P(A)$  area of the cyan rectangle



Tossing a dice .....

2D1431 Machine Learning

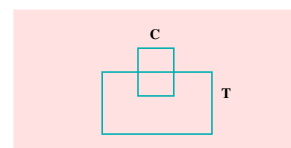
Danica Kragic



## Conditional Probability

$P(A|B)$  - fraction of worlds in which  $B$  is true that also have  $A$  true

- $T$  - have a toothache
  - $C$  - have a cavity
- $$P(T) = 1/10$$
- $$P(C) = 1/30$$
- $$P(T|C) = 1/2$$
- Toothache is rare and cavity is even rarer, but if you already have a cavity, there is a 50-50 chance that you will get toothache. Note! Not all cavities give toothache.



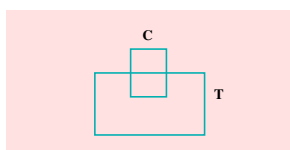
2D1431 Machine Learning

Danica Kragic

## Conditional Probability

- $P(T|C)$  fraction of “cavity” worlds in which you also have a toothache

$$\begin{aligned}
 &= \frac{\text{\#worlds with cavity and tootache}}{\text{\#worlds with cavity}} \\
 &= \frac{\text{\#Area of C and T}}{\text{\#Area of C}} \\
 &= \frac{P(C, T)}{P(C)}
 \end{aligned}$$



2D1431 Machine Learning

Danica Kragic

## Bayes Theorem

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- $P(h)$  = prior probability of hypothesis  $h$  - PRIOR
- $P(D)$  = prior probability of training data  $D$  - EVIDENCE
- $P(h|D)$  = probability of  $h$  given  $D$  - POSTERIOR
- $P(D|h)$  = probability of  $D$  given  $h$  - LIKELIHOOD

2D1431 Machine Learning

Danica Kragic

## Bayes Theorem

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

posterior = likelihood  $\times$  prior / evidence

- By observing the data  $D$  we can convert the prior probability  $P(h)$  to the a posteriori probability  $P(h|D)$
- The posterior is the probability that  $h$  holds after data  $D$  has been observed
- The evidence  $P(D)$  can be viewed merely as a scale factor that guarantees that the posterior probabilities sum to 1

2D1431 Machine Learning

Danica Kragic

## Bayes Theorem

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- Goal: To determine the most probable hypothesis, given the data  $D$  plus any initial knowledge about the prior probabilities of the various hypotheses in  $H$ .
- Prior probability of  $h$ ,  $P(h)$ : it reflects any background knowledge we have about the chance that  $h$  is a correct hypothesis (before having observed the data).
- Prior probability of  $D$ ,  $P(D)$ : it reflects the probability that training data  $D$  will be observed given no knowledge about which hypothesis  $h$  holds.
- Conditional Probability of observation  $D$ ,  $P(D|h)$ : it denotes the probability of observing data  $D$  given some world in which hypothesis  $h$  holds.

2D1431 Machine Learning

Danica Kragic

## Bayes Theorem, Cont.

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

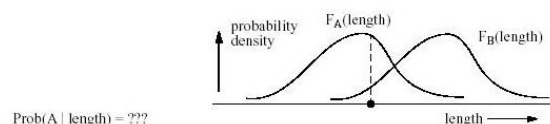
- Posterior probability of  $h$ ,  $P(h|D)$ : it represents the probability that  $h$  holds given the observed training data  $D$ . It reflects our confidence that  $h$  holds after we have seen the training data  $D$  and it is the quantity that Machine Learning researchers are interested in.
- Bayes Theorem allows us to compute  $P(h|D)$ :

2D1431 Machine Learning

Danica Kragic

## Another example

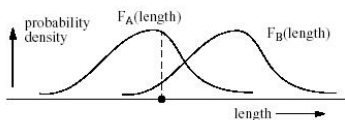
- Given: a couple of classes of objects (men, women) distributions over parameters (hair length, height, weight)
- Task: given a new object *person*, with known hair length, which class does it belong to?



2D1431 Machine Learning

Danica Kragic

### What if we are in a boys' private school?



Prob(A | length) = ???

$$\text{Prob}(A | \text{length}) \text{Prob}(\text{length}) = \text{Prob}(A, \text{length}) = \text{Prob}(\text{length} | A) \text{Prob}(A)$$

$$\text{Prob}(A | \text{length}) = \frac{\text{Prob}(\text{length} | A) \text{Prob}(A)}{\text{Prob}(\text{length})} = \frac{\text{Prob}(\text{length} | A) \text{Prob}(A)}{\text{Prob}(\text{length} | A) \text{Prob}(A) + \text{Prob}(\text{length} | B) \text{Prob}(B)}$$

$$\text{Prob}(A | \text{length}) = \frac{F_A(\text{length}) P_A}{F_A(\text{length}) P_A + F_B(\text{length}) P_B}$$

### Terminology

- **Maximum A Posteriori (MAP) and Maximum Likelihood (ML) Hypotheses**
  - MAP hypotheses: Highest conditional probability given observations (data)
  - ML: highest likelihood of generating the observed data
- **Bayesian Inference:** Computing Conditional Probabilities in a Model
- **Bayesian Learning:** Searching Model (Hypothesis) Space using Conditional Probabilities

### Choosing Hypotheses - MAP

$$P(h|D) = \frac{P(D|h) P(h)}{P(D)}$$

In general, we want the most probable hypothesis given the data  
*Maximum a posteriori* hypothesis  $h_{MAP}$ :

$$\begin{aligned} h_{MAP} &= \arg \max_{h \in H} P(h|D) \\ &= \arg \max_{h \in H} \frac{P(D|h)P(h)}{P(D)} \\ &= \arg \max_{h \in H} P(D|h)P(h) \end{aligned}$$

### Choosing Hypotheses - ML

If we assume

$$P(h_i) = P(h_j)$$

we can further simplify and choose the *Maximum likelihood* (ML) hypothesis

$$h_{ML} = \arg \max_{h_i \in H} P(D|h_i)$$

### An example: Does a patient have cancer or not?

A patient takes a lab test and the result comes back positive. The test returns a correct positive result in only 98% of the cases in which the disease is actually present, and a correct negative result in only 97% of the cases in which the disease is not present. Furthermore, .008 of the entire population have cancer.

Hypotheses  $h$ :            disease,             $\neg$  disease  
 Priors  $P(h)$ :         $P(\text{disease}) = 0.008$      $P(\neg \text{disease}) = 0.992$   
 Likelihoods  $P(D|h)$ :     $P(+ | \text{disease}) = 0.98$      $P(- | \text{disease}) = 0.02$   
                                   $P(+ | \neg \text{disease}) = 0.03$      $P(- | \neg \text{disease}) = 0.97$

### Cont.

Maximum posteriors  $\arg \max P(h|D)$ :

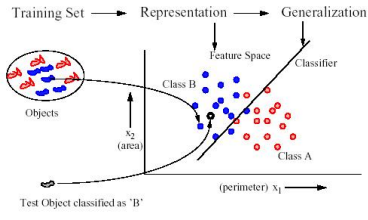
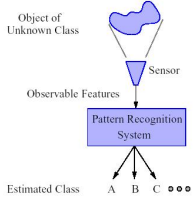
$$P(\text{dis} | +) \sim P(+ | \text{dis})P(\text{dis}) = 0.0078$$

$$P(\neg \text{dis} | +) \sim P(+ | \neg \text{dis})P(\neg \text{dis}) = 0.0298$$

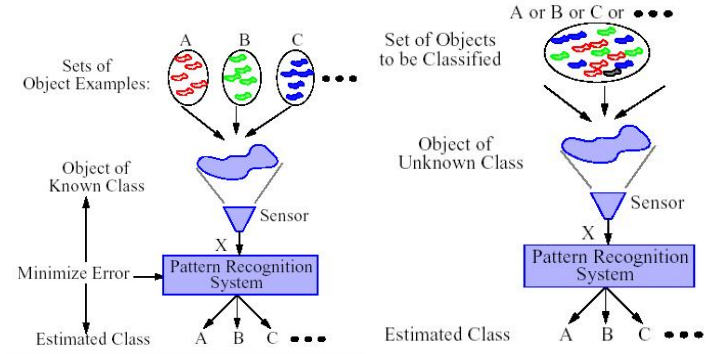
$$P(\text{dis} | +) = 0.0078 / (0.0078 + 0.298) = 0.21$$

$$P(\neg \text{dis} | +) = 0.0298 / (0.0078 + 0.298) = 0.79$$

### Computer Vision

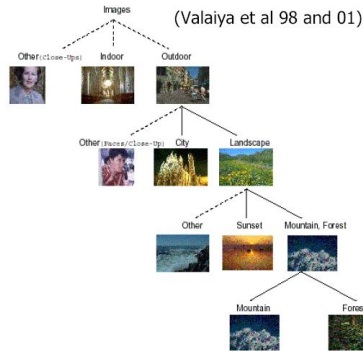
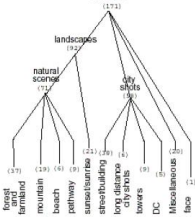


### Statistical Pattern Recognition Learning

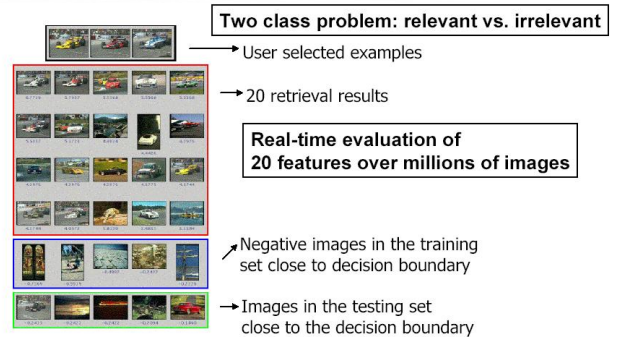


### Example

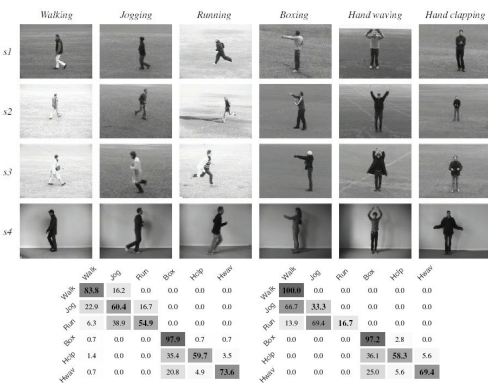
- How to select the categories and tree?
- How to estimate the distributions of features for each class?



### Example



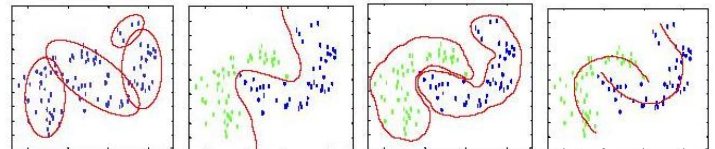
### Laptev et al 2004



### Pattern recognition

From left to right:

- Clustering:** find natural groups of samples in unlabeled data
- Classification:** find functions separating the classes
- Density estimation:** make a statistical model of the data
- Regression:** fit lines or other functions to data



## Bayesian Learning - Features

- A training example can increase/decrease the probability of a correct hypothesis (no elimination)
- Prior knowledge incorporated
- Accommodate hypotheses with probabilistic predictions
- Classification of new instances based on combined predictions of multiple hypotheses
- Provide standard for optimal decision making

## Relation to Pattern Classification

- given a set of measurements represented with a pattern vector  $x$  (instance)
- assign the pattern to one of  $H$  classes  $h_i$
- **Decision rule** partitions the measurement space in  $H$  regions  $\Omega_i$
- if an observation vector is in  $\Omega_i$ , it is assumed to belong to class  $h_i$
- The boundaries between regions  $\Omega_i$ : **decision boundaries** or **decision surfaces**

## Bayes Decision Rule: Two class problem

- Only priors known:  $P(h_j) > P(h_k) \quad k = 1 \dots H; \quad k \neq j$
- Given observation data:

$$P(h_j|D) > P(h_k|D) \quad k = 1 \dots H; \quad k \neq j$$

- or **Bayes rule for minimum error**

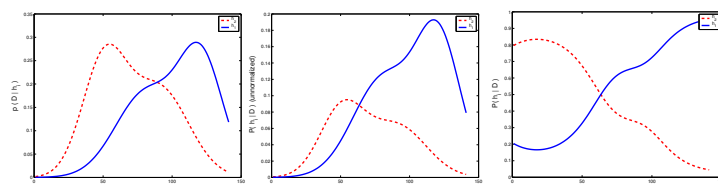
$$P(D|h_j)P(h_j) > P(D|h_k)P(h_k)$$

- Decision rule (likelihood ratio):

$$l_r(D) = \frac{P(D|h_j)}{P(D|h_k)} > \frac{P(h_k)}{P(h_j)} \text{ implies } D \in \text{class } h_j$$

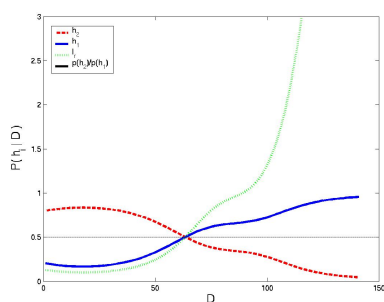
## Example

- two classes,  $p(h_1) = 2/3$  and  $p(h_2) = 1/3$
- using Bayes rule



## Example, cont.

Estimating decision rule given likelihood ratio:



## Discriminant functions

- $f(D)$  that leads to a classification rule

$$f(D) > k \rightarrow D \in h_1$$

$$f(D) < k \rightarrow D \in h_2$$

- instead of making assumptions about  $p(D|h_i)$ , we make assumptions about the form of  $f(D)$
- Two class problem

$$f(D) = \frac{P(D|h_1)}{P(D|h_2)}$$

with  $k = P(h_2)/P(h_1)$

### Discriminant functions

- In general, decision surface between classes  $i$  and  $j$  for minimum error classification can be defined as

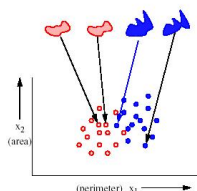
$$P(h_i|D) - P(h_j|D) = 0$$

- On one side the difference is positive, on the other negative.
- Instead of working directly with probabilities, it may be more convenient to use another equivalent function

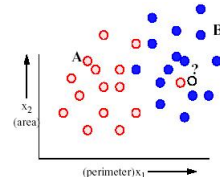
$$g_i(D) = f(P(h_i|D))$$

where  $f$  is any monotonically increasing function

### Problems .....



Similar objects are close in feature space; Different objects may be close or remote!!

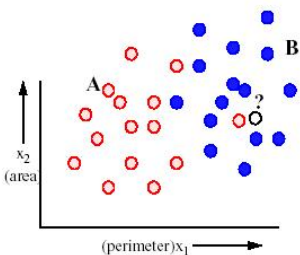


? to be classified as

A - because it is most closest to an object A

B - because the local density of B is larger

### Estimation .....



What is the probability of finding an object of class A (B) on this place in the 2D space?



What is the probability of finding an object of class A (B) on this place in the 1D space?

### Normal density

- Multivariate normal (aka Gaussian) density adequate model for many applications (also easy to analyze).
- Univariate (one variable) normal pdf

$$N(\mu, \sigma^2) \sim p(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

with mean  $\mu$  and variance  $\sigma^2$  ( $\sigma$  is the standard deviation).

- Multivariate I-dimensional Gaussian pdf

$$N(\mu, \Sigma) \sim p(x) = \frac{1}{(2\pi)^{l/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right]$$

where  $\mu = E[x]$  is the mean vector, and  $\Sigma$  is the  $l \times l$  covariance matrix

$$\Sigma = E[(x-\mu)(x-\mu)^T]$$

### Normal density, Cont.

- If the  $i$ -th and  $j$ -th dimensions are statistically or linearly independent then  $E(x_i x_j) = E(x_i)E(x_j)$  and  $\sigma_{ij} = 0$
- If all dimenions are statistically or linearly independent, then  $\sigma_{ij} = 0, \forall i \neq j$  and  $\Sigma$  has non-zero elements only on the diagonal
- If the underlying density is Gaussian and  $\Sigma$  is a diagonal matrix, then the dimensions are statistically independent and

$$p(x) = \prod_i p(x_i)$$

$$p(x_i) = N(\mu_i, \sigma_{ii})$$

### Bayes Decision Rule

Assume that we have two Gaussian distributions associated to two separate classes  $c_1$  and  $c_2$

$$P(x|c_i) = P(x|\mu_i, \sigma_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\sum_i (x_i - \mu_i)^2 / 2\sigma_i^2\right]$$

Bayes decision rule (max posterior probability)

Decide  $c_1$  if  $P(c_1|x) > P(c_2|x)$

otherwise decide  $c_2$

if  $P(c_1) = P(c_2)$  use maximum likelihood  $P(x|c_i)$

else use maximum posterior  $P(c_i|x) = P(x|c_i)P(c_i)$

### Two category case

Discriminant functions: if  $g(x) > 0$  then  $c_1$  else  $c_2$

$$g(x) = P(c_1|x) - P(c_2|x)$$

$$= P(x|c_1)P(c_1) - P(x|c_2)P(c_2)$$

$$g(x) = \ln P(c_1|x) - \ln P(c_2|x)$$

$$= \ln P(x|c_1)/P(x|c_2) - \ln P(c_1)/P(c_2)$$

Gaussian probability functions with identical  $\Sigma_i$

$$g(x) = -\frac{(x-\mu_1)^2}{2\sigma^2} + \frac{(x-\mu_2)^2}{2\sigma^2} + \ln P(c_1) - \ln P(c_2)$$

decision surface is a line/hyperplane

### Two class case

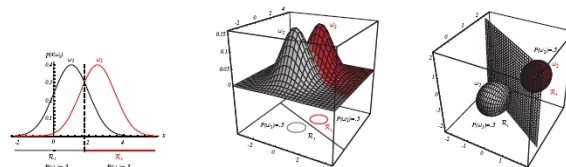


FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in  $d$  dimensions, and the boundary is a generalized hyperplane of  $d - 1$  dimensions, perpendicular to the line separating the means. In these one-, two-, and three-dimensional examples, we indicate  $p(x|\omega_i)$  and the boundaries for the case  $P(\omega_1) = P(\omega_2)$ . In the three-dimensional case, the grid plane separates  $\mathcal{R}_1$  from  $\mathcal{R}_2$ . From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

### Two category case

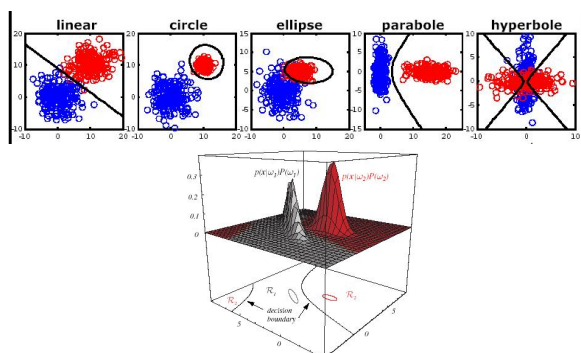


FIGURE 2.6. In this two-dimensional two-category classifier the probability densities are Gaussian, the decision boundary consists of two hyperbolas, and thus the decision region  $\mathcal{R}_2$  is not simply connected. The ellipses mark where the density is  $1/e$  times that at the peak of the distribution. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

### Basic Formulas for Probabilities

- **Product Rule:** probability  $P(A \wedge B)$  of a conjunction of two events A and B:

$$P(A \wedge B) = P(A|B)P(B) = P(B|A)P(A)$$

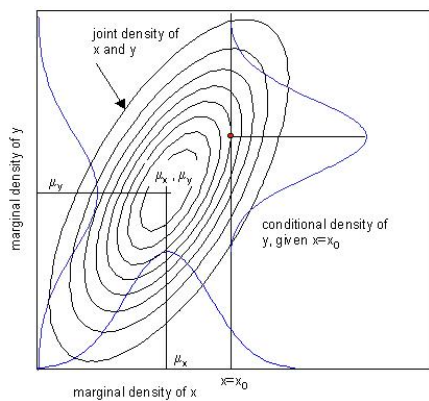
- **Sum Rule:** probability of a disjunction of two events A and B:

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

- **Theorem of total probability:** if events  $A_1, \dots, A_n$  are mutually exclusive with  $\sum_{i=1}^n P(A_i) = 1$ , then

$$P(B) = \sum_{i=1}^n P(B|A_i)P(A_i)$$

### Joint, marginal, and conditional densities



### Brute Force MAP Hypothesis Learner

1. For each hypothesis  $h$  in  $H$ , calculate the posterior probability

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

2. Output the hypothesis  $h_{MAP}$  with the highest posterior probability

$$h_{MAP} = \operatorname{argmax}_{h \in H} P(h|D)$$

## Relation to Concept Learning

Consider our usual concept learning task

- instance space  $X$ , hypothesis space  $H$ , training examples  $D$
- consider the `FINDS` learning algorithm: Given  $D$ , outputs most specific hypothesis  $h$  in the version space  $VS_{H,D}$

What would Bayes rule produce as the MAP hypothesis?

Does `FindS` output a MAP hypothesis??

## Relation to Concept Learning

Assume: Fixed set of instances  $(x_1, \dots, x_m)$  and  $D$  is the set of classifications  $D = \langle c(x_1), \dots, c(x_m) \rangle$   
Choose  $P(D|h)$

- $P(D|h) = 1$  if  $h$  consistent with  $D$
- $P(D|h) = 0$  otherwise

Choose  $P(h)$  to be *uniform* distribution

- $P(h) = \frac{1}{|H|}$  for all  $h$  in  $H$

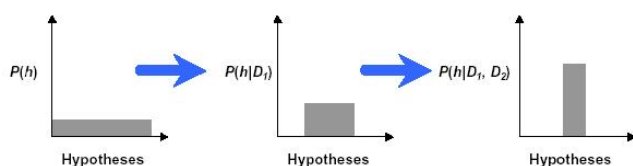
Then,

$$P(h|D) = \begin{cases} \frac{1}{|VS_{H,D}|} & \text{if } h \text{ is consistent with } D \\ 0 & \text{otherwise} \end{cases}$$

- All hypotheses start with with same probability then the inconsistent hypotheses drop to zero while the rest share equally.
- Every consistent hypothesis is a MAP hypothesis.

## Evolution of Posterior Probabilities

- Start with uniform priors (equal probabilities assigned to each hypothesis)



- Evidential inference

- Introduce data (evidence)  $D_1$ : belief revision occurs (Learning agent revises conditional probability of inconsistent hypotheses to 0; Posterior probabilities for remaining  $h \in VS_{H,D}$  revised)
- Add more data (evidence)  $D_2$ : further belief revision

## Parameter Estimation

So far,

- we have assumed that conditional probabilities  $p(D|h)$  and priors  $P(h), P(D)$  were known
- This is a dream!
- Have to estimate these from training data
- Problem when  $D$  size significant

## Parameter Estimation, Cont.

- Assumptions help!
- From unknown function  $p(D|h)$  to parameter estimation  $\mu, \sigma$
- Maximum-likelihood estimation
- Bayesian estimation

## Maximum Likelihood vs. Bayesian Estimation

- ML: parameters are quantities with unknown but fixed values
- ML: the best estimate maximizes the probability of obtaining the observed samples
- BE: parameters are random variables with some prior distribution
- BE: observation of samples converts these to posterior density, revising our belief about the true values of the parameters



## Maximum Likelihood

Recall:

$$h_{ML} = \arg \max_h P(D|h)$$

- MLE: the parameter value that maximizes the likelihood function
- If we assume that  $D$  contains  $n$  samples  $x_1, \dots, x_n$  (independence assumption)

$$P(D|h) = \prod_{k=1}^n P(x_k|h)$$

- Often we work with "log" of the function (*log-likelihood*,  $l(h)$ )

$$l(h) = \ln p(D|h)$$

$$h_{ML} = \arg \max_h l(h) = \arg \max_h \sum_{k=1}^n \ln P(x_k|h)$$

- Now, calculate the first derivative of  $l(h)$ , set it to 0 and solve the equation for  $h$ !

2D1431 Machine Learning

Danica Kragic



## Maximum Likelihood: Example

Assume normal distribution with unknown mean and variance:

$$p(x_k|h) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_k - \mu)^2}{2\sigma^2}\right)$$

$$h(h_1 = \mu, h_2 = \sigma^2)$$

$$\ln p(x_k|h) = -\frac{1}{2} \ln 2\pi h_2 - \frac{1}{2h_2} (x_k - h_1)^2$$

derivative .....

solution .....

2D1431 Machine Learning

Danica Kragic



## Summary

- Bayesian theory: combines prior knowledge and observed data to find the most likely hypotheses
- MAP and ML hypotheses
- Bayesian theory and Pattern Recognition: a tool for estimating decision functions
- Parameter estimation: ML and Bayesian estimation

2D1431 Machine Learning

Danica Kragic