Machine Learning

Chapter 6 Bayesian Learning

6.1 Introduction

- Why introduce Bayesian learning into machine learning
  - Bayesian reasoning provides a probabilistic approach to inference
    - calculate explicit probabilities for hypotheses,
      - such as the naïve Bayes classifier
  - Provide basis for analyzing many learning algorithms that do not explicitly manipulate probabilities
    - Find-S
    - Candidate elimination algorithms
    - Neural network
    - Inductive bias of decision tree
    - Minimum Description Length principle

Features of Bayesian learning

- Each training examples can change the probability that a hypotheses is correct
- The final probability of a hypotheses depends on Prior knowledge and observed data
  - Prior knowledge:
    - prior probability P(h)
    - a probability distribution over observed data for each h, P(D/h)
- Allow hypotheses make probabilistic predictions
- Combining the predictions of multiple hypotheses for classification.
  - Weighted by their probabilities
- A standard of optimal decision making

Difficulty in applying Bayesian methods

- Requires initial knowledge of many probabilities.
  - If they are not known, they should be estimated

- The significant computational cost to find the Bayes optimal hypothesis

Content

- Introduce Bayes theorem
- Define maximum likelihood and maximum a posteriori probability hypotheses
- Analyze several issues and learning algorithms
- Introduce several learning algorithms that explicitly manipulate probabilities
  - Bayes optimal classifier
  - Gibbs algorithm
  - Naïve Bayes classifier
  - Bayes belief networks

6.2 Bayes Theorem

- Task of machine learning: find the best hypothesis, given the observed training data D.
  - The best hypothesis: the most probable hypothesis given the data D plus the knowledge about the prior probabilities of hypotheses (Prior knowledge)
    - Bayes theorem provides a way to calculate the probability of a hypotheses based on training data and Prior knowledge .

Prior probability and posterior probability

- P(h) (prior probability of h): the probability h holds, before we have observed the training data.
  - reflect any background knowledge about the chance that h is a correct hypothesis
  - No such prior knowledge, assums the same prior probability to each h

- P(D) : the prior probability that training data D will be observed
- P(D|h) : the probability of observing D given some world in which h holds
- P(h|D) (posterior probability): the probability that h holds given the observed training data D.

Bayes theorem

- Provide a way to calculate the posterior probability P(h|D) from the prior probability P(h)

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

  - P(h|D) increases with P(h) and with P(D|h)
  - decreases as P(D) increases.
    - Why?

Maximum a posteriori hypothesis

- Learner finds the most probable hypothesis h given the observed data D.
  - maximum a posteriori (MAP) hypothesis

$$h_{MAP} = \arg\max_{h \in H} P(h \mid D)$$

$$= \arg\max_{h \in H} \frac{P(D \mid h) P(h)}{P(D)}$$

$$= \arg\max_{h \in H} P(D \mid h) P(h)$$

Maximum likelihood hypothesis

- Assume that every hypothesis has the same priori ($P(h_i)=P(h_j)$).
  - P(D|h) is called the likelihood of the data D given h

- Any h that maximizes P(D|h) is called a maximum likelihood hypothesis

$$h_{ML} = \arg\max_{h \in H} P(D \mid h)$$

6.2.1 An Example

- Two hypotheses: patient has cancer, patient does not
- Laboratory test with two possible outcomes: + and -

- Prior knowledge: only 0.008 have this disease.
- The test returns a correct positive result in only 98% of the cases in which the disease is present
- The test returns a correct negative result in only 97% of the cases in which the disease is not present

- In summary

  $P(cancer)=0.008$, $P(\neg cancer)=0.992$

  $P(+|cancer)=0.98$, $P(-|cancer)=0.02$

  $P(+|\neg cancer)=0.03$, $P(-|\neg cancer)=0.97$

- problem： A patient for whom the test returns a positive result. The patient has cancer or not?

- Find MAP hypothesis $\arg\max_{h \in H} P(D \mid h)P(h)$
  - $P(h_1=cancer|+)$： $P(+|cancer)P(cancer)=0.0078$
  - $P(h_2=\neg\ cancer|+)$： $P(+|\neg cancer)P(\neg cancer)=0.0298$
  - $h_{MAP}=h_2=\neg cancer$
    - P(D)?

- Bayesian inference depends strongly on the prior probabilities.
- hypotheses are not completely accepted or rejected,

Summary of basic probability formulas

- Product rule： $P(A \wedge B)=P(A|B)P(B)=P(B|A)P(A)$
- Sum rule： $P(A \vee B)=P(A)+P(B)-P(A \wedge B)$
- Bayes theorem： $P(h|D)=P(D|h)P(h)/P(D)$
- Theorem of total probability：
  - If events $A_1 \ldots A_n$ are mutually exclusive with

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

6.3 Bayes Theorem and Concept Learning

- Brute-force Bayesian concept learning algorithm
- Analyzing the concept learning algorithm in chapter 2,
  - outputs MAP hypotheses

6.3.1 Brute-Force Bayes Concept Learning

- Concept learning problem: for finite hypothesis Space H, the task is to learn target concept c:X->(0,1)

- Brute-Force MAP Learning algorithm
  - For each h in H, calculate the posterior probability

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

  - Output the h_{MAP}

$$h_{MAP} = \arg\max_{h \in H} P(h \mid D)$$

  - Require significant computation.
  - Impractical for large hypothesis spaces, but it provides a standard against which we may judge the performance of other concept learning algorithms

Special settings for Brute-force MAP learning algorithm

- Assumptions:
  - The training data D is noise free (i.e., $d_i=c(x_i)$)
  - c is contained in H
  - Every h has the same probability to occur

- So we have

$$P(h)=\frac{1}{|H|}$$

$$P(D|h)=\begin{cases}1 & \forall d_i, d_i=h(x_i)\\ 0 & otherwise\end{cases}$$

- The first step of Brute-Force MAP
  - h is inconsistent with D， $P(h \mid D) = \dfrac{0 \cdot P(h)}{P(D)} = 0$
  - h is consistent with D， $P(h \mid D) = \dfrac{1 \cdot \frac{1}{|H|}}{P(D)} = \dfrac{\frac{1}{|H|}}{\frac{|VS_{H,D}|}{|H|}} = \dfrac{1}{|VS_{H,D}|}$

$VS_{H,D}$ is the subset of hypotheses from H that are consistent with D

- Derive P(D) from the theorem of total probability

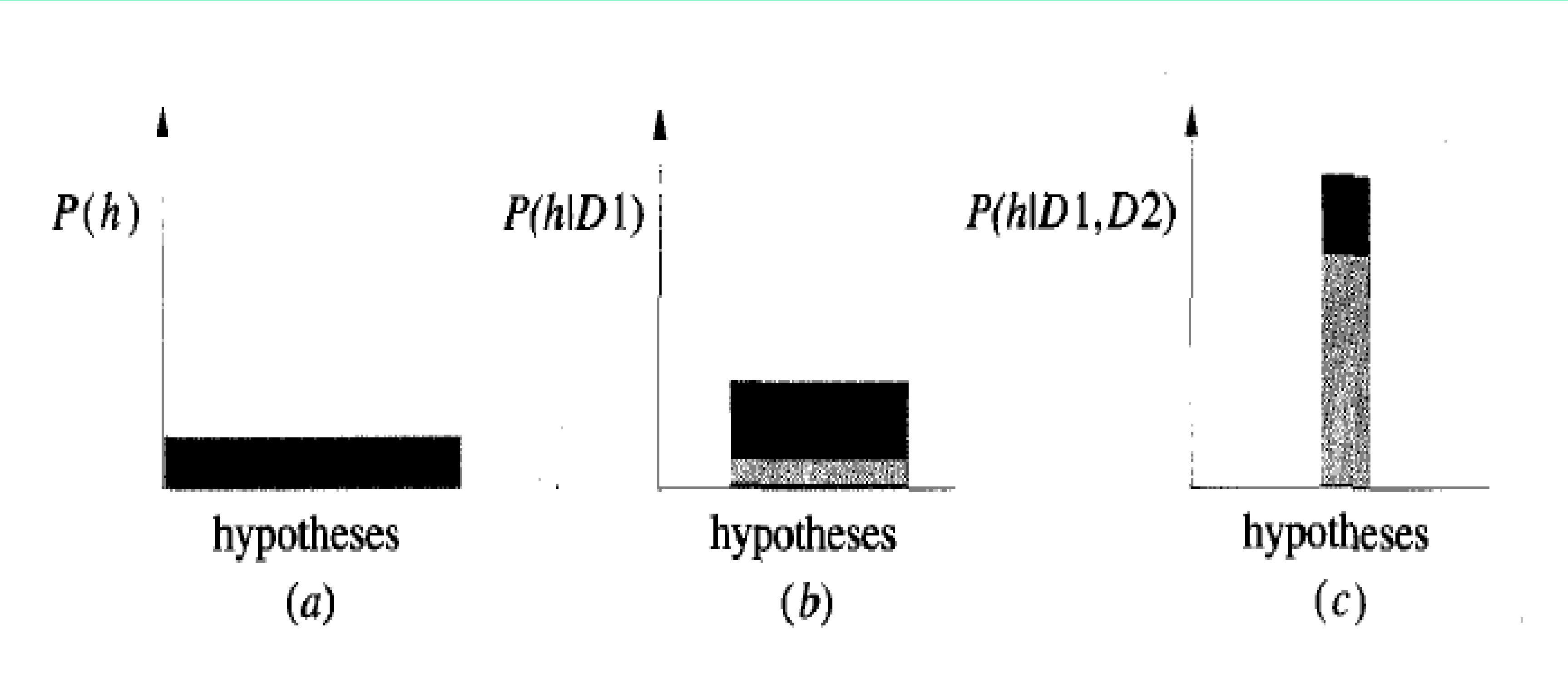
P(D) $$= \sum_{h_i \in H} P(D \mid h_i) P(h_i)$$

$$= \sum_{h_i \in VS_{H,D}} 1 \times \frac{1}{|H|} + \sum_{h_i \notin VS_{H,D}} 0 \times \frac{1}{|H|}$$

$$= \sum_{h_i \in VS_{H,D}} 1 \times \frac{1}{|H|}$$

$$= \frac{|VS_{H,D}|}{|H|}$$

- The evolution of probability is depicted in fig 6.1. (Next page)



Conclusion：Every consistent hypothesis is a MAP hypothesis

6.3.2 MAP Hypotheses and Consistent Learners

- Consistent learners:
  - a learning algorithm outputs a hypothesis that commits zero errors over the training examples

- Every consistent learner outputs a MAP hypothesis
  - assume a uniform prior probability distribution over H
  - assume deterministic, noise-free training data

- Find-S
  - output a MAP hypothesis under the above probability distributions
  - outputs MAP hypothesis that favors more specific hypotheses
    - $p(h_i)>p(h_j)$ if h_i is more specific than h_j.

- The Bayesian framework provides one way to characterize the behavior of learning algorithms
  - Define P(h) and P(D|h) under which the algorithm outputs optimal hypotheses
    - P(h) and P(D|h) can been seen as the implicit assumptions under which this algorithm behaves optimally
      - P(h), the prior probability over H
      - P(D|h), the strength of data in rejecting or accepting a hypothesis
    - In fact, P(h) and P(D|h) are also the inductive bias of the learner
      - In chapter 2 we define the inductive bias to be the set of assumptions B sufficient to deductively justify the inductive inference.

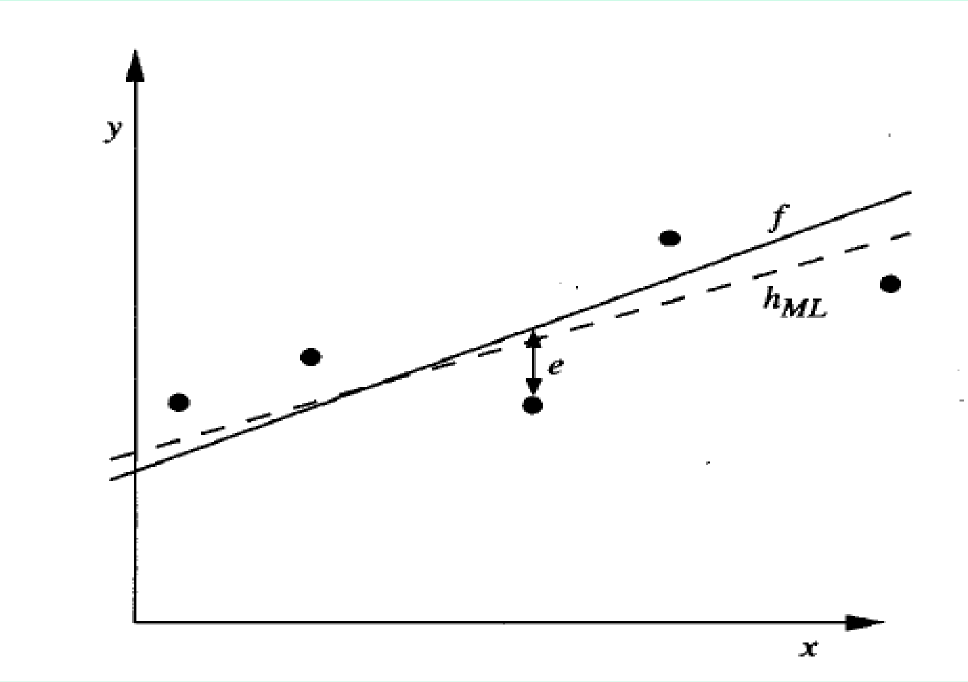
– By defining P(h) and P(D|h), the inductive inference method can be modeled as an equivalent probabilistic reasoning based on Bayes theorem
  - A probability reasoning system based on Bayes theorem will exhibit input-output behavior equivalent to these algorithms

6.4 Maximum likelihood and least-squared error hypotheses

- A Bayesian analysis shows that
  - under certain assumptions, any learning algorithm that minimizes the squared error will output a maximum likelihood hypothesis
  - The significance
    - provides a Bayes-based interpretation for the outputs of ANN and other curve fitting methods

- Problem setting：
  - Learner L (X, H, D)
    - an instance space X
    - a hypothesis space H (real-valued functions)
    - D: m training examples, where the target value is corrupted by noise which follows Normal distribution
      - $<x_i,d_i>$， where $d_i=f(x_i)+e_i$
      - $f(x_i)$: true value
      - e_i is random noise and drawn independently from a normal distribution with 0 mean
  - The task of learner
    - output a maximum likelihood hypothesis
    - or, equivalently, a MAP hypothesis (assume all hypotheses are equally probable a priori)

– A simple example, a linear function Fig(6-2)



– Define of Probability density function：

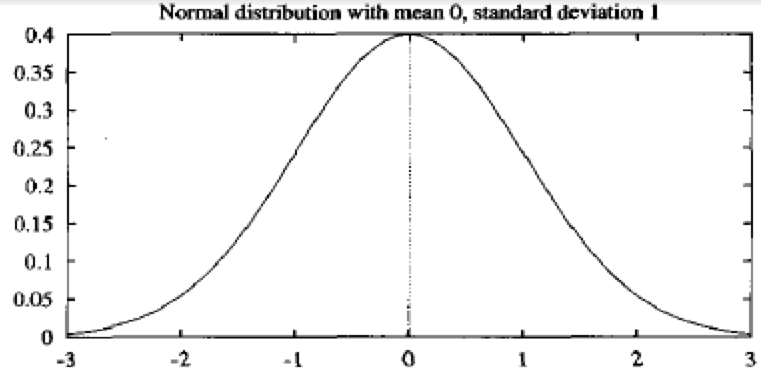
$$p(x_0) = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} P(x_0 \le x < x_0 + \varepsilon)$$

Prove

- Define h_{ML}
  - Assuming the training examples are mutually independent
  - $P(D|h)$=the product of the various $p(d_i|h)$

$$h_{ML} = \arg\max_{h \in H} \prod_{i=1}^{m} p(d_i \mid h)$$

- Define Data distribution
  - noise e_i obeys $N(0, \sigma)$
  - d_i must also obey $N(f(x_i), \sigma)$
  - $p(d_i|h)$ can be written as $N(f(x_i), \sigma)$ (table 5-4, Next Page)



A Normal distribution (also called a Gaussian distribution) is a bell-shaped distribution defined by the probability density function

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

A Normal distribution is fully determined by two parameters in the above formula: μ and σ.

If the random variable X follows a normal distribution, then:

- The probability that X will fall into the interval (a, b) is given by

$$\int_a^b p(x)dx$$

- The expected, or mean value of X, $E[X]$, is

$$E[X] = \mu$$

- The variance of X, $Var(X)$, is

$$Var(X) = \sigma^2$$

- The standard deviation of X, σ_X, is

$$\sigma_X = \sigma$$

The Central Limit Theorem (Section 5.4.1) states that the sum of a large number of independent, identically distributed random variables follows a distribution that is approximately Normal.

- substitute $\mu=f(x_i)=h(x_i)$,

we have:

$$h_{ML} = \arg\max_{h\in H} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(d_i-\mu)^2}$$

$$= \arg\max_{h\in H} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(d_i-h(x_i))^2}$$

$$= \arg\max_{h\in H} \sum_{i=1}^{m} \ln \frac{1}{\sqrt{2\pi\sigma^2}} -- \frac{1}{2\sigma^2}(d_i - h(x_i))^2$$

$$= \arg\max_{h\in H} \sum_{i=1}^{m} -\frac{1}{2\sigma^2}(d_i - h(x_i))^2$$

$$= \arg\min_{h\in H} \sum_{i=1}^{m} (d_i - h(x_i))^2$$

- Proves that the maximum likelihood hypothesis is the one that minimizes the sum of the squared errors between d_i and $h(x_i)$
  - Constraints:
    - d_i are generated by adding random noise to the true target value
    - this random noise obeys a normal distribution with zero mean

Why choose the normal distribution to characterize noise?

- A mathematically straightforward analysis

- It is a good approximation to many types of noise in physical system

- The central limit theorem
  - this implies that noise generated by the sum of very many independent, but identically distributed factors will itself be normally distributed.