Foundations of Machine Learning

Module 4:
Part B: Bayesian Learning

Sudeshna Sarkar
IIT Kharagpur

Probability for Learning

- Probability for classification and modeling concepts.
- Bayesian probability
  - Notion of probability interpreted as partial belief
- Bayesian Estimation
  - Calculate the validity of a proposition
    - Based on prior estimate of its probability
    - and New relevant evidence

Bayes Theorem

- **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*.

Bayes Theorem

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

- P(h) = prior probability of hypothesis h
- P(D) = prior probability of training data D
- P(h|D) = probability of h given D (posterior density)
- P(D|h) = probability of D given h (likelihood of D given h)

An Example

Does 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 this cancer.

$$P(cancer) = .008, P(\neg cancer) = .992$$

$$P(+ \mid cancer) = .98, P(- \mid cancer) = .02$$

$$P(+ \mid \neg cancer) = .03, P(- \mid \neg cancer) = .97$$

$$P(cancer \mid +) = \frac{P(+ \mid cancer)P(cancer)}{P(+)}$$

$$P(\neg cancer \mid +) = \frac{P(+ \mid \neg cancer)P(\neg cancer)}{P(+)}$$

Maximum A Posteriori (MAP) Hypothesis

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

The Goal of Bayesian Learning: the most probable hypothesis given the training data (Maximum A Posteriori 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 (ML) Hypothesis

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

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

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

- If every hypothesis in *H* is equally probable a priori, we only need to consider the likelihood of the data *D* given *h*, ***P(D|h).*** Then, h_{MAP} becomes the ***Maximum Likelihood***,

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

MAP Learner

For each hypothesis h in H, calculate the posterior probability

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

Output the hypothesis h_{MAP} with the highest posterior probability

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

Comments:

Computational intensive

Providing a standard for judging the performance of learning algorithms

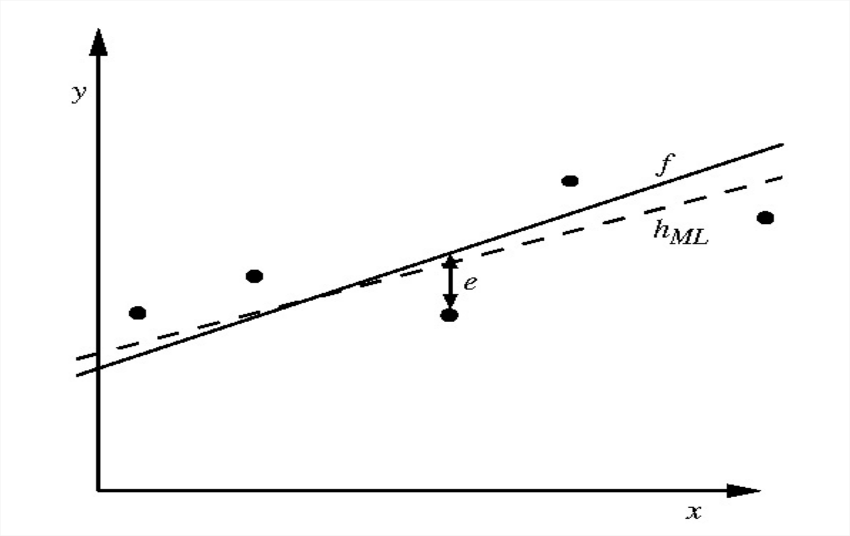
Choosing P(h) and P(D|h) reflects our prior knowledge about the learning task

Maximum likelihood and least-squared error

- Learn a Real-Valued Function:
  - Consider any real-valued target function f.
  - Training examples (x_i,d_i) are assumed to have Normally distributed noise e_i with zero mean and variance σ^2, added to the true target value f(x_i),

d_i satisfies $N(f(x_i), \sigma^2)$

Assume that e_i is drawn independently for each x_i .



Compute ML Hypo

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

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

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

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

Bayes Optimal Classifier

Question: Given new instance x, what is its most probable classification?

- $h_{MAP}(x)$ is not the most probable classification!

Example: Let P(h1|D) = .4, P(h2|D) = .3, P(h3 |D) =.3

Given new data x, we have h1(x)=+, h2(x) = -, h3(x) = -

What is the most probable classification of x ?

Bayes optimal classification:

$$\arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j \mid h_i) P(h_i \mid D)$$

where *V* is the set of all the values a classification can take and v_j is one possible such classification.

Example:

P(h1| D) =.4,	P(-|h1)=0,	P(+|h1)=1	$\sum_{hi \in H} P(+ \mid h_i) P(h_i \mid D) = .4$
P(h2|D) =.3,	P(-|h2)=1,	P(+|h2)=0	
P(h3|D)=.3,	P(-|h3)=1,	P(+|h3)=0	$\sum_{hi \in H} P(- \mid h_i) P(h_i \mid D) = .6$

Why "Optimal"?

- Optimal in the sense that no other classifier using the same *H* and prior knowledge can outperform it on average

Gibbs Algorithm

- Bayes optimal classifier is quite computationally expensive, if *H* contains a large number of hypotheses.
- An alternative, less optimal classifier Gibbs algorithm, defined as follows:
  1. Choose a hypothesis randomly according to $P(h|D)$, where *D* is the posterior probability distribution over *H*.
  2. Use it to classify new instance

Error for Gibbs Algorithm

- Surprising fact: Assume the expected value is taken over target concepts drawn at random, according to the prior probability distribution assumed by the learner, then (Haussler *et al.* 1994)

$$E_f[error_{X,f}GibbsClassifier] \leq 2E_f[error_{X,f}BayesOPtimal],$$

where f denotes a target function, X denotes the instance space.

Thank You