BAYESIAN THEOREM

- Bayes TheoremMAP,
- ML hypothesesMAP learners
- Minimum description length principle
- Bayes optimal classifier
- Naïve Bayes learner
- Bayesian belief networks

**Two Roles for Bayesian Methods**

Provide practical learning algorithms:

- Naïve Bayes learningBayesian belief network learning
- Combine prior knowledge (prior probabilities) with observed data

Requires prior probabilities:

- Provides useful conceptual framework:
- Provides “gold standard” for evaluating other learning algorithms
- Additional insight into Occam’s razor
- Bayes Theorem
- P(h) = prior probability of hypothesis h
- P(D) = prior probability of training data D
- P(h|D) = probability of h given D
- P(D|h) = probability of D given h

- Choosing Hypotheses
- Generally want the most probable hypothesis given the training data
- Maximum a posteriori hypothesis hMAP:
- If we assume P(hi)=P(hj) then can further simplify, and choose the Maximum likelihood (ML) hypothesis
- Bayes Theorem
- 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, 0.8% of the entire population have this cancer.

- P(cancer) = P( cancer) =
- P(+|cancer) = P(-|cancer) =
- P(+| cancer) = P(-| cancer) =
- P(cancer|+) =P( cancer|+) =

**Some Formulas for Probabilities**

Product rule:

probability P(A B) of a conjunction of two events A and B:

P(A B) = P(A|B)P(B) = P(B|A)P(A)

Sum rule:

probability of disjunction of two events A and B:

P(A B) = P(A) + P(B) - P(A B)

Theorem of total probability: if events A1,…,An are mutually exclusive with , then

***Brute Force MAP Hypothesis Learner***

- 1. For each hypothesis h in H, calculate the posterior probability
- 2. Output the hypothesis hMAP with the highest posterior probability

**Bayes Optimal Classifier**

- Bayes optimal classification
- Example:P(h1|D)=.4,
- P(-|h1)=0,
- P(+|h1)=1P(h2|D)=.3,
- P(-|h2)=1,
- P(+|h2)=0P(h3|D)=.3,
- P(-|h3)=1,
- P(+|h3)=0therefore

**Gibbs Classifier**

- Bayes optimal classifier provides best result, but can be expensive if many hypotheses.
- Gibbs algorithm:
- 1. Choose one hypothesis at random, according to P(h|D)
- 2. Use this to classify new instanceSurprising fact: assume target concepts are drawn at random from H according to priors on H.
- Then:E[errorGibbs] 2E[errorBayesOptimal]
- Suppose correct, uniform prior distribution over H, thenPick any hypothesis from VS, with uniform probabilityIts expected error no worse than twice Bayes optimal

***Naïve Bayes Classifier***

- Along with decision trees, neural networks, nearest neighor, one of the most practical learning methods.
- When to useModerate or large training set availableAttributes that describe instances are conditionally independent given classification
- Successful applications:DiagnosisClassifying text documents

**Summary of Bayes Belief Networks**

- Combine prior knowledge with observed dataImpact of prior knowledge (when correct!) is to lower the sample complexity
- Active research areaExtend from Boolean to real-valued variablesParameterized distributions instead of tables
- Extend to first-order instead of propositional systems
- More effective inference methods