Bayesian Learning

• It involves direct manipulation of probabilities in order to find correct hypotheses

• The quantities of interest are governed by probability distributions

• Optimal decisions can be made by reasoning about those probabilities
Bayesian Learning

• Bayesian learning algorithms are among the most practical approaches to certain type of learning problems

• They provide a useful perspective for understanding many learning algorithms that do not explicitly manipulate probabilities
Features of Bayesian Learning

- Each training example can incrementally decrease or increase the estimated probability that a hypothesis is correct.

- Prior knowledge can be combined with observed data to determine the final probability of a hypothesis.

- Hypotheses with probabilities can be accommodated.

- New instances can be classified by combining multiple hypotheses weighted by their probabilities.
Bayes Theorem

\[
P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}
\]

- \( P(h) \): prior probability of hypothesis \( h \)
- \( P(D) \): prior probability of training data \( D \)
- \( P(h \mid D) \): probability that \( h \) holds given \( D \)
- \( P(D \mid h) \): probability that \( D \) is observed given \( h \)
Bayes Theorem

- Maximum *A-posteriori* hypothesis (MAP):
  (dependent on experience)

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

*P(h)* is not a uniform distribution over *H*. 
Bayes Theorem

• Maximum Likelihood hypothesis (ML):

\[ h_{ML} = \arg\max_{h \in H} P(h | D) = \arg\max_{h \in H} P(D | h) \]

\[ P(h) \text{ is a uniform distribution over } H. \]
Bayes Theorem

- 0.008 of the population have cancer
- Positive results are classified correctly in only 98% cases
- Negative results are classified correctly in only 97% cases

Would a new patient with a positive result have cancer or not?
Bayes Theorem

• 0.008 of the population have cancer

• Positive results are classified correctly in only 98% cases

• Negative results are classified correctly in only 97% cases

Would a new patient with a positive result have cancer or not?

\[ P(\text{cancer} | \oplus) \leq P(\neg \text{cancer} | \oplus) \]
Bayes Theorem

- $P(\text{cancer}) = .008 \Rightarrow P(\neg \text{cancer}) = .992$
- $P(\oplus | \text{cancer}) = .98$
- $P(\ominus | \neg \text{cancer}) = .97 \Rightarrow P(\oplus | \neg \text{cancer}) = .03$

$$P(\text{cancer} | \oplus) \approx P(\oplus | \text{cancer})P(\text{cancer}) = .0078$$
$$P(\neg \text{cancer} | \oplus) \approx P(\oplus | \neg \text{cancer})P(\neg \text{cancer}) = .0298$$
Bayes Theorem

- Maximum *A-posteriori* hypothesis (MAP):

\[
h_{\text{MAP}} = \arg \max_{h \in \{\text{cancer}, \neg \text{cancer}\}} P(h | \oplus)
\]

\[
= \arg \max_{h \in \{\text{cancer}, \neg \text{cancer}\}} P(\oplus | h)P(h)
\]

\[
= \neg \text{cancer}
\]
Bayes Theorem and Concept Learning

\[
P(h | D) = \frac{P(D | h)P(h)}{P(D)}
\]

- \( P(h) = 1/|H| \)
- \( P(D | h) = 1 \) iff \( h \) is consistent with \( D \); 0 otherwise
- \( P(D) = |\text{VersionSpace}_{HD}|/|H| \)
Bayes Theorem and Concept Learning

FIND-S:

• $P(h_1) \geq P(h_2)$ if $h_1$ is more specific than $h_2$

$\Rightarrow h_{MAP} = \underset{h \in H}{\text{argmax}} \ P(D \mid h)P(h) = h_{\text{most specific}}$
Maximum Likelihood and Least-Squared Error Hypotheses
Maximum Likelihood and Least-Squared Error Hypotheses

\[ h_{ML} = \arg\max_{h \in H} P(D | h) \]

\[ = \arg\max_{h \in H} \prod_{i=1,m} P(d_i | h) \]

\[ = \arg\max_{h \in H} \prod_{i=1,m} k.e^{-c(d_i - h(x_i))^2} \]

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

Normal distribution
Minimum Description Length Principle

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

\[ = \arg\max_{h \in H} \log_2 P(D \mid h) + \log_2 P(h) \]

\[ = \arg\min_{h \in H} - \log_2 P(D \mid h) - \log_2 P(h) \]

• \(- \log_2 P(h)\): the description length of \( h \) under optimal encoding

• \(- \log_2 P(D \mid h)\): the description of \( D \) given \( h \) under optimal encoding
Bayes Optimal Classifier

• What is the most probable hypothesis given the training data?

• What is the most probable classification of a new instance given the training data?
Bayes Optimal Classifier

• Hypothesis space = \{h_1, h_2, h_3\}

• Posterior probabilities = \{.4, .3.\, .3\} (h_1 is h_{\text{MAP}})

• New instance \(x\) is classified positive by \(h_1\) and negative by \(h_2\) and \(h_3\)

What is the most probable classification of \(x\)?
Bayes Optimal Classifier

• The most probable classification of a new instance is obtained by combining the predictions of all hypotheses weighted by their posterior probabilities:

\[
\text{argmax}_{c \in C} P(c \mid D) = \text{argmax}_{c \in C} \sum_{h \in H} P(c \mid h).P(h \mid D)
\]
Naive Bayes Classifier

- Each instance $x$ is described by a conjunction of attribute values $<a_1, a_2, ..., a_n>$

- The target function $f(x)$ can take on any value from a finite set $C$

- It is to assign the most probable target value to a new instance
Naive Bayes Classifier

\[ c_{\text{MAP}} = \arg \max_{c \in C} P(c | a_1, a_2, ..., a_n) \]
\[ = \arg \max_{c \in C} P(a_1, a_2, ..., a_n | c).P(c) \]

\[ c_{\text{NB}} = \arg \max_{c \in C} \prod_{i=1,n} P(a_i | c).P(c) \]

assuming that \( a_1, a_2, ..., a_n \) are independent given \( c \)
Naive Bayes Classifier

\[ x = (\text{Outlook} = \text{sunny}, \text{Temp} = \text{cool}, \text{Humidity} = \text{high}, \text{Wind} = \text{strong}) \]

- \( P(\text{Play Tennis} = \text{yes} \mid x) = ? \)
- \( P(\text{Play Tennis} = \text{no} \mid x) = ? \)
Naive Bayes Classifier

\[ x = (\text{Outlook} = \text{sunny}, \text{Temp} = \text{cool}, \text{Humidity} = \text{high}, \text{Wind} = \text{strong}) \]

- \( P(\text{Play Tennis} = \text{yes} \mid x) = ? \)
- \( P(\text{Play Tennis} = \text{no} \mid x) = ? \)

\[ c \left( a_1, a_2, ..., a_n \right) \]
Naive Bayes Classifier

Estimating probabilities:

\[
\frac{n_c + mp}{n + m}
\]

- \(n\): total number of training examples of a particular class
- \(n_c\): number of training examples having a particular attribute value in that class
- \(m\): equivalent sample size
- \(p\): prior estimate of the probability (= \(1/k\) where \(k\) is the number of possible values of the attribute)
Naive Bayes Classifier

Learning to classify text:

\[ c_{NB} = \text{argmax}_{c \in C} \prod_{i=1,n} P(a_i = w_k \mid c).P(c) \]

\[ = \text{argmax}_{c \in C} \prod_{i=1,n} P(w_k \mid c).P(c) \]

assuming that all words have equal chance occurring in every position
Exercises

- In Mitchell’s ML (Chapter 6): 6.1 to 6.4