

Bayesian Learning

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)

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

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

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

Bayes Theorem

- Maximum **Likelihood** hypothesis (ML):

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

$P(h)$ is a uniform distribution over H .

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

Bayes Theorem

- 0.008 of the population have cancer
- Only 98% patients are correctly classified as positive
- Only 97% non-patients are correctly classified as negative

Would a person with a positive result have cancer or not?

Bayes Theorem

- 0.008 of the population have cancer
- Only 98% patients are correctly classified as positive
- Only 97% non-patients are correctly classified as negative

Would a person with a positive result have cancer or not?

$$P(\text{cancer}|\oplus) \geq < P(\neg\text{cancer}|\oplus) ?$$

Bayes Theorem

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

$$h_{\text{MAP}} = \operatorname{argmax}_{h \in \{\text{cancer}, \neg \text{cancer}\}} P(h \mid \oplus)$$

$$= \operatorname{argmax}_{h \in \{\text{cancer}, \neg \text{cancer}\}} P(\oplus \mid h)P(h)$$

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):

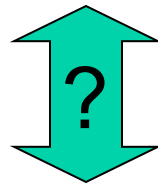
$$h_{\text{MAP}} = \operatorname{argmax}_{h \in \{\text{cancer}, \neg\text{cancer}\}} P(h \mid \oplus)$$

$$= \operatorname{argmax}_{h \in \{\text{cancer}, \neg\text{cancer}\}} P(\oplus \mid h)P(h)$$

$$= \neg\text{cancer}$$

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_{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**:

$$\operatorname{argmax}_{c \in C} P(c \mid D)$$

$$= \operatorname{argmax}_{c \in C} \sum_{h \in H} P(c \mid h) \cdot P(h \mid D)$$

Naive Bayes Classifier

- Each instance x is described by a conjunction of attribute values $\langle a_1, a_2, \dots, a_n \rangle$
- 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

$$\begin{aligned}c_{\text{MAP}} &= \operatorname{argmax}_{c \in C} P(c \mid a_1, a_2, \dots, a_n) \\ &= \operatorname{argmax}_{c \in C} P(a_1, a_2, \dots, a_n \mid c).P(c)\end{aligned}$$

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

assuming that a_1, a_2, \dots, a_n are independent given c

Naive Bayes Classifier

Example	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cold	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes
5	Cloudy	Warm	High	Weak	Cool	Same	Yes
6	Cloudy	Cold	High	Weak	Cool	Same	No

7	Sunny	Warm	Normal	Strong	Warm	Same	?
8	Sunny	Warm	Low	Strong	Cool	Same	?

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:

$$\begin{aligned} c_{NB} &= \operatorname{argmax}_{c \in C} \prod_{i=1, n} P(a_i = w_k | c) \cdot P(c) \\ &= \operatorname{argmax}_{c \in C} \prod_{i=1, n} P(w_k | c) \cdot P(c) \end{aligned}$$

position *i* in the text

assuming that all words have equal chance occurring in every position

Exercises

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