# Bayesian Learning

# Basics of Probability

- $P(A)$: probability that $A$ happens
- $P(A|B)$: probability that $A$ happens, given that $B$ happens (conditional probability)
- Some rules:
  - Complement: $P(A^C) = 1 - P(A)$
  - Disjunction: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  - Conjunction: $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$
  - If $A$ and $B$ are independent, $P(A \cap B) = P(A)P(B)$
  - Total probability: $P(B) = \sum_{i=1}^{k} P(A|B_i)P(B_i)$

# Conditional Probability

- If:  the patient has symptom of toothache
- Then: conclude cavity with probability $P$
- where $P$ is the following conditional probability

$$P(\text{cavity}|\text{toothache})$$

- To compute p(cavity|toothache), we can compute

$$p(\text{cavity} \wedge \text{toothache}) / p(\text{cavity})$$

# Bayes' Rule

- By definition of conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B|A)}{P(B)}$$

  - P(A): prior probability of hypothesis A
  - P(A|B): posterior probability of A given evidence B
  - P(B|A): likelihood of B given A
  - P(B): prior probability of B

# Example: Conditional Probability

- T+ = toothache (positive test)
- T- = normal (negative test)
- C+ = cavity (disease present)
- C- = no cavity (disease absent)

| | C+ | C- | Total |
|---|---|---|---|
| T+ | 25 | 14 | 39 |
| T- | 18 | 78 | 96 |
| Total | 43 | 92 | 135 |

# Example: Conditional Probability

- T+ = toothache (positive test)
- T- = normal (negative test)
- C+ = cavity (disease present)
- C- = no cavity (disease absent)

| | C+ | C- | Total |
|---|---|---|---|
| T+ | 0.185 | 0.104 | 0.289 |
| T- | 0.133 | 0.578 | 0.711 |
| Total | 0.318 | 0.682 | 1 |

# Example: Conditional Probability

- T+ = toothache (positive test)
- T- = normal (negative test)
- C+ = cavity (disease present)
- C- = no cavity (disease absent)

If toothache, then conclude cavity with probability p(C+|T+)=?

| | C+ | C- | Total |
|---|---|---|---|
| T+ | 0.185 | 0.104 | 0.289 |
| T- | 0.133 | 0.578 | 0.711 |
| Total | 0.318 | 0.682 | 1 |

# Argmax/Argmin

argmax stands for the argument of the maximum, that is to say, the set of points of the given argument for which the given function attains its maximum value.

$$\operatorname*{argmax}_x f(x) = \{x | \forall y : f(y) \leq f(x)\}$$

$$\operatorname*{argmax}_x (-|x|) = \{0\}$$

# Maximum A Posteriori

Find the most probable hypothesis given the training data (Maximum A Posteriori hypothesis $H_{map}$)

$$
\begin{aligned}
H_{MAP} &= \text{argmax}_{H \in \mathbb{H}} P(H|E) \\
&= \text{argmax}_{H \in \mathbb{H}} \frac{P(H)P(E|H)}{P(E)} \\
&= \text{argmax}_{H \in \mathbb{H}} P(H)P(E|H)
\end{aligned}
$$

# Maximum Likelihood

Sometimes we may assume that all a priori probabilities are equally likely in which case the method is called a **maximum likelihood** or ML method

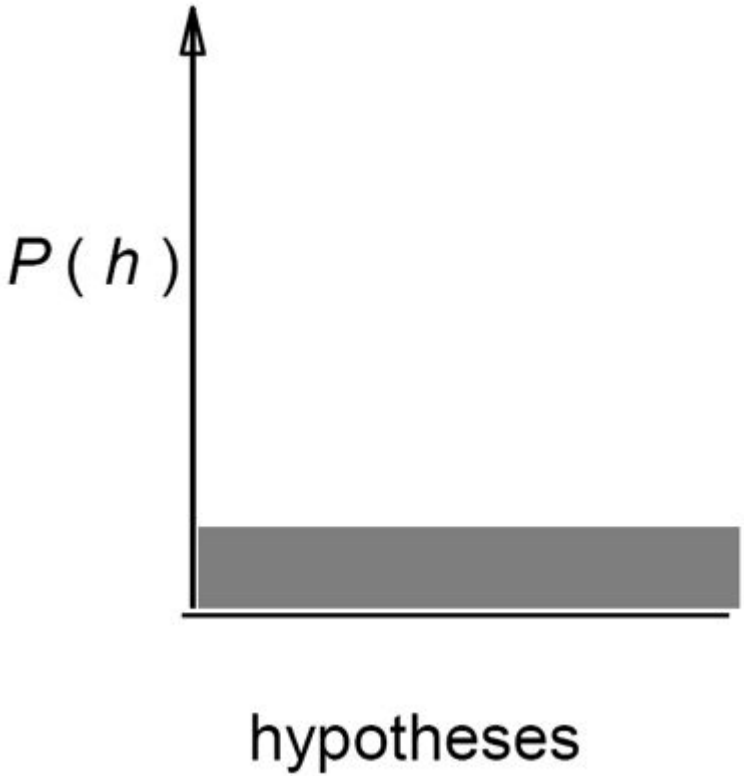
$$H_{ML} = \text{argmax}_{H \in \mathbb{H}} P(E|H)$$

# Example

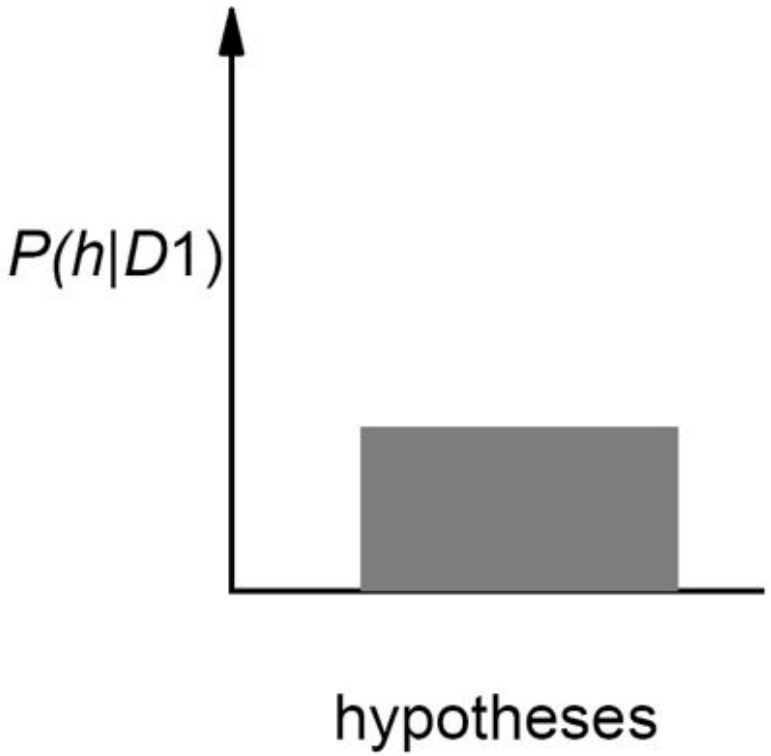
A patient takes a lab test and the result comes back positive. The test returns a correct positive result in only 98% of the cases in which the disease is actually present, and a correct negative result in only 97% of the cases in which the disease is not present. Also, 0.008 of the entire population have this cancer.

For a new patient that lab test returns a positive result, should he be diagnosed as having cancer or not?

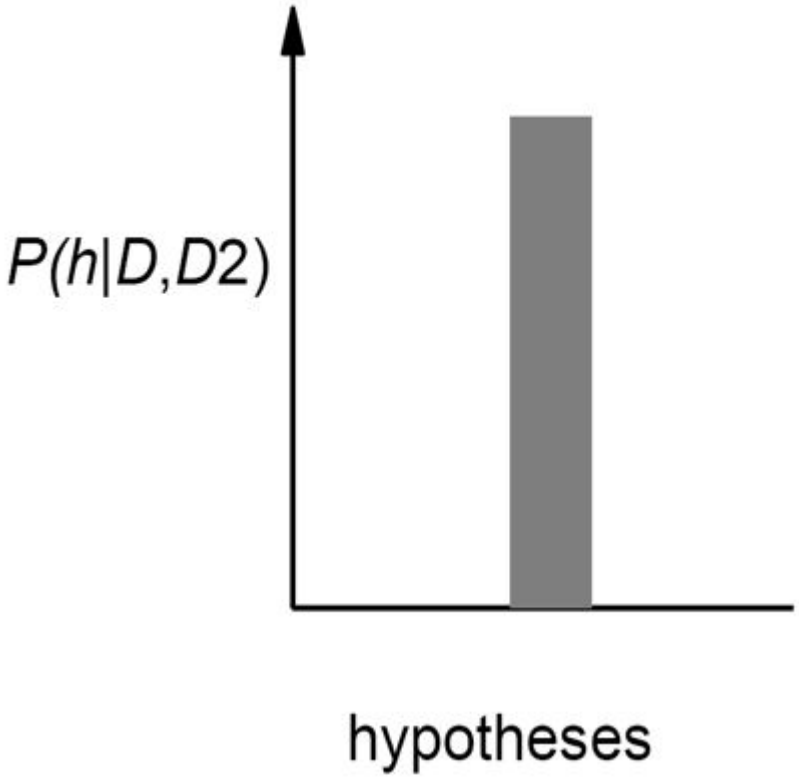
# Evolution of Posterior Probabilities



# Evolution of Posterior Probabilities



# Evolution of Posterior Probabilities



# Classification Using Bayes Rule

Given multiple attribute values, what is the most probable value of the target variable?

$$
\begin{aligned}
h_{MAP} &= \underset{h_i \in \mathbb{H}}{\operatorname{argmax}}\ p(h_i | d_1, d_2, \cdots, d_n) \\
&= \underset{h_i \in \mathbb{H}}{\operatorname{argmax}}\ \frac{p(h_i) p(d_1, d_2, \cdots, d_n | h_i)}{p(d_1, d_2, \cdots, d_n)} \\
&= \underset{h_i \in \mathbb{H}}{\operatorname{argmax}}\ p(h_i) p(d_1, d_2, \cdots, d_n | h_i)
\end{aligned}
$$

Problem: too much data needed to estimate $p(d_1,d_2,\ldots,d_n|h_i)$

# Classification Using Bayes Rule

- What is the usual way to compute $p(d_1,d_2,\ldots,d_n|h_i)$?

- Chain rule:

$$p(d_1,d_2,\ldots,d_n|h_i) = p(d_1|h_i) \times p(d_2|h_i,d_1) \times p(d_3|h_i,d_1,d_2) \times \ldots \times p(d_n|h_i,d_1,d_2,\ldots,d_{n-1})$$

# Naïve Bayes Classifier

# Naïve Bayes Classifier

- Based on Bayes' rule + assumption of conditional independence
  - assumption often violated in practice
  - even then, it usually works well

- Successful application: classification of text documents

# Conditional Independence

- $p(d_1,d_2,\ldots,d_n|h_i) = p(d_1|h_i) \times p(d_2|h_i,d_1) \times p(d_3|h_i,d_1,d_2) \times \ldots \times p(d_n|h_i,d_1,d_2,\ldots,d_{n-1})$

- **Naïve Bayes** (conditionally independence) assumption : attributes are independent, given the class
  - $p(d_2|h_i,d_1)=p(d_2|h_i)$
  - $p(d_3|h_i,d_1,d_2)=p(d_3|h_i)$

    …,
  - $p(d_n|h_i,d_1,d_2,\ldots,d_{n-1})=p(d_n|h_i)$

# Naïve Bayes Classifier

$$
\begin{aligned}
h_{NB} &= \underset{h_i \in \mathbb{H}}{\operatorname{argmax}}\ p(h_i | d_1, d_2, \cdots, d_n) \\
&= \underset{h_i \in \mathbb{H}}{\operatorname{argmax}}\ \frac{p(h_i)p(d_1, d_2, \cdots, d_n | h_i)}{p(d_1, d_2, \cdots, d_n)} \\
&= \underset{h_i \in \mathbb{H}}{\operatorname{argmax}}\ p(h_i)p(d_1, d_2, \cdots, d_n | h_i) \\
&= \underset{h_i \in \mathbb{H}}{\operatorname{argmax}}\ p(h_i) \prod_{j=1}^{n} p(d_j | h_i)
\end{aligned}
$$

# Naïve Bayes Classifier

- What if the independence assumption is violated, that is,

$$p(d_1,d_2,\ldots,d_n|h_i) \neq p(d_1|h_i)\times p(d_2|h_i) \times p(d_3|h_i) \times \ldots \times p(d_n|h_i)?$$

- Prediction is still equivalent to Bayes prediction as long as the following weaker condition holds:

$$\begin{aligned}&\operatorname*{argmax}_{h_i \in \mathbb{H}} p(d_1, d_2, \cdots, d_n|h_i)p(h_i)\\ &= \operatorname*{argmax}_{h_i \in \mathbb{H}} p(d_1|h_i)p(d_2|h_i)\cdots p(d_n|h_i)p(h_i)\end{aligned}$$

# Learning a Naïve Bayes Classifier

- We need to estimate $p(h_i)$, $p(d_1|h_i)$, $p(d_2|h_i)$,…,$p(d_n|h_i)$ from data.
- Then $h_{NB} = \underset{h_i \in \mathbb{H}}{\operatorname{argmax}}\, p(h_i) \prod_{j=1}^{n} p(d_j|h_i)$

- How to estimate?
  - Simplest: standard estimate from statistics
    - estimate probability from sample proportion
    - e.g., estimate $p(A|B)$ as $count\ (A\ and\ B)$ / $count(B)$

# Naïve Bayes Example

| Day | Outlook | Temp. | Humidity | Wind | Play (Tennis) |
|---|---|---|---|---|---|
| D1 | Sunny | Hot | High | Weak | No |
| D2 | Sunny | Hot | High | Strong | No |
| D3 | Overcast | Hot | High | Weak | Yes |
| D4 | Rain | Mild | High | Weak | Yes |
| D5 | Rain | Cool | Normal | Weak | Yes |
| D6 | Rain | Cool | Normal | Strong | No |
| D7 | Overcast | Cool | Normal | Strong | Yes |
| D8 | Sunny | Mild | High | Weak | No |
| D9 | Sunny | Cold | Normal | Weak | Yes |
| D10 | Rain | Mild | Normal | Weak | Yes |
| D11 | Sunny | Mild | Normal | Strong | Yes |
| D12 | Overcast | Mild | High | Strong | Yes |
| D13 | Overcast | Hot | Normal | Weak | Yes |
| D14 | Rain | Mild | High | Strong | No |

A New Day: Outlook= sunny, Temperature=cool, humidity=high and windy= strong, play tennis?

# Naïve Bayes Example

Consider a medical diagnosis problem with three possible diagnoses (well, cold, allergy, ) and three symptoms (sneeze, cough, fever)

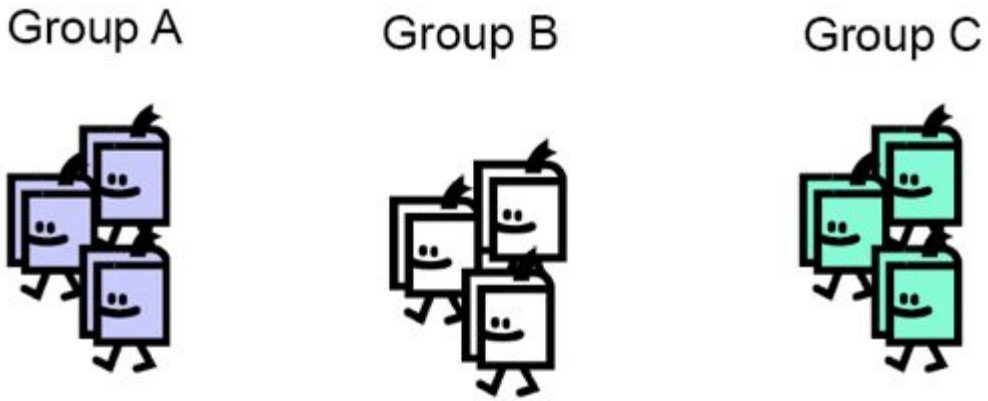
| Diagnoses | Well | Cold | Allergy | Rhinitis |
|---|---|---|---|---|
| p(h) | 0.8 | 0.05 | 0.05 | 0.1 |
| p(sneeze\|h) | 0.1 | 0.9 | 0.9 | 0.8 |
| p(cough\|h) | 0.1 | 0.8 | 0.7 | 0.6 |
| p(fever\|h) | 0.01 | 0.7 | 0.4 | 0.6 |

If diagnosis is sneeze, cough and ¬fever, what is the conclusion -- well, cold or allergy ?

# Text Classification using Naïve Bayes

# Text Classification

- Given text of newsgroup article, guess which newsgroup it is taken from.
- Naïve Bayes turns out to work well on this application.
- Key issue : how do we represent examples? what are the attributes?



# Text Classification

- Class $h_j$: Binary classification (+/−) or multiple classes possible $H$ ($j$ = 1,2, …, $k$)

- How about attributes?

# Example

- 1000 training documents that someone has 700 classified as “dislikes” ($h_0$) and 300 classified as “likes” ($h_1$).
- Suppose document 1 is “**This is a very interesting document**”

$$h_{NB} = \max_{h_j \in \{\text{like,dislike}\}} p(h_j) \times p(d_1 = \text{this}|h_j) \times$$
$$p(d_2 = \text{is}|h_j) \cdots \times p(d_6 = \text{document}|h_j)$$

$p(like)$=300/1000=0.3

$p(dislike)$=1−$p(like)$=0.7

- How to estimate $p(d_i|h_j)$?

# Comments on Naïve Bayes Learner

- One of the most practical learning methods, along with decision trees, neural networks, etc.
- Requires :
  - Moderate or large training data set
  - Attributes that describe instances should be conditionally independent given the classification.
- Successful applications include diagnosis and text classification.
- It may not estimate probabilities accurately when independence is violated, it still picks correct category