Bayesian Learning

Bayesian Classification: Why?

- A statistical classifier: performs *probabilistic prediction*, *i.e.*, predicts class membership probabilities
- Foundation: Based on Bayes' Theorem.
- Performance: A simple Bayesian classifier, *naïve Bayesian classifier*, has comparable performance with decision tree and selected neural network classifiers
- Incremental: Each training example can incrementally increase/decrease the probability that a hypothesis is correct — prior knowledge can be combined with observed data
- Standard: Even when Bayesian methods are computationally intractable, they can provide a standard of optimal decision making against which other methods can be measured

Things We'd Like to Do

- Spam Classification
  - Given an email, predict whether it is spam or not
- Medical Diagnosis
  - Given a list of symptoms, predict whether a patient has disease X or not
- Weather
  - Based on temperature, humidity, etc… predict if it will rain tomorrow

Probability Basics

- Prior, conditional and joint probability for random variables
  - Prior probability: $P(x)$
  - Conditional probability: $P(x_1 | x_2), P(x_2 | x_1)$
  - Joint probability: $\mathbf{x} = (x_1, x_2), P(\mathbf{x}) = P(x_1, x_2)$
  - Relationship: $P(x_1, x_2) = P(x_2 | x_1)P(x_1) = P(x_1 | x_2)P(x_2)$
  - Independence:

    $P(x_2 | x_1) = P(x_2), P(x_1 | x_2) = P(x_1), P(x_1, x_2) = P(x_1)P(x_2)$

Variables

- Boolean random variables:
  - cavity might be true or false
- Discrete random variables: weather might be sunny, rainy, cloudy, snow
  - $P(Weather=sunny)$
  - $P(Weather=rainy)$
  - $P(Weather=cloudy)$
  - $P(Weather=snow)$
- Continuous random variables:
  - the temperature has continuous values

Basic Concepts

- **Sample space**: the set of possible outcomes of an experiment.
- **Event**: a subset of the sample space.
- If S is a finite sample space of equally likely outcomes, and E is an event, that is, a subset of S, then the *probability* of E is

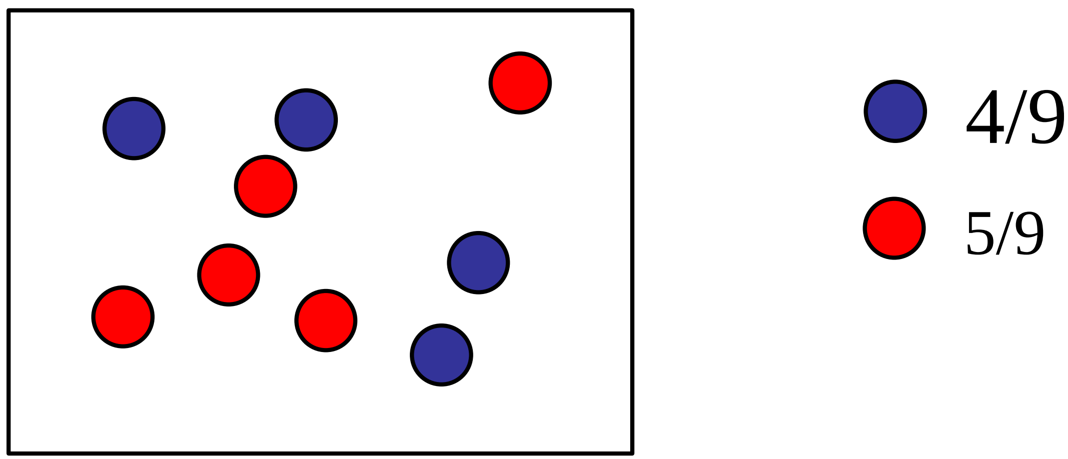
$$p(E) = \frac{|E|}{|S|}$$

Probability

- Before the evidence is obtained; prior probability
  - *P(a):* the prior probability that the proposition is true
  - $P(cavity)=0.1$

- After the evidence is obtained; posterior probability
  - $P(a|b)$
  - The probability of *a* given that all we know is *b*
  - $P(cavity|toothache)=0.8$

Examples



Axioms of Probability

- All probabilities are between 0 and 1. For any proposition *a*,

$$0 \leq P(a) \leq 1$$

- The probability of disjunction is given by

$$P(a \cup b) = P(a) + P(b) - P(a \cap b)$$

- Product rule

$$P(a \cap b) = P(a \mid b)P(b)$$

$$P(a \cap b) = P(b \mid a)P(a)$$

Theorem of total probability

If events A_1, ... , A_n are mutually

exclusive with $\sum_{i=1}^{n} P(A_i) = 1$ then

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

$$P(B) = \sum_{i=1}^{n} P(B, A_i)$$

Bayes's rule

$$P(b \mid a) = \frac{P(a \mid b)P(b)}{P(a)}$$

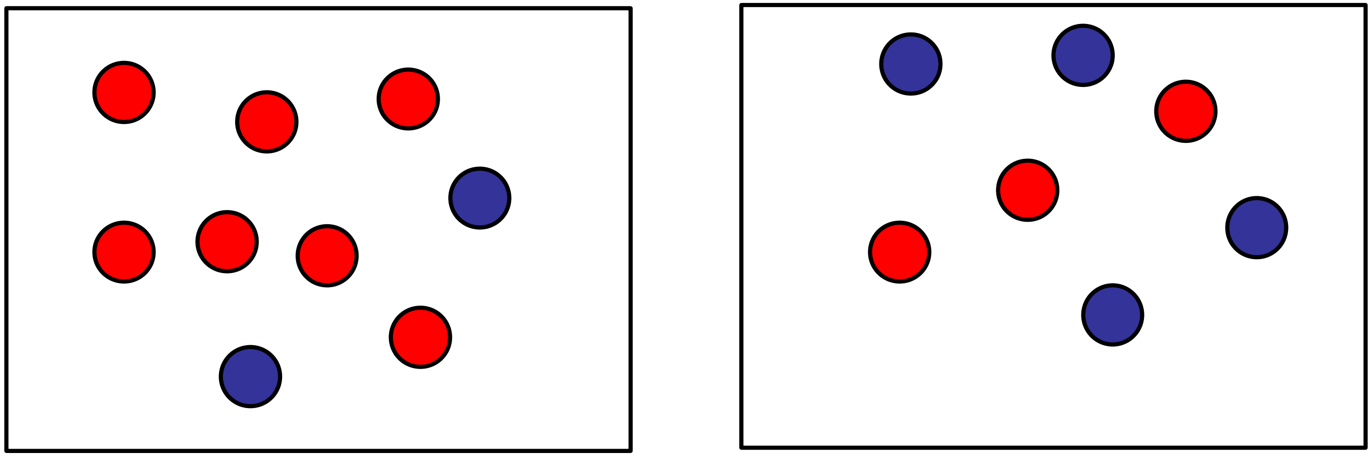
$$P(c \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid c)P(c)}{P(\mathbf{x})}$$

$$Posterior = \frac{Likelihood \times Prior}{Evidence}$$

Discriminative

Generative

Example



1. Choose one of the two boxes at random.
2. Select one of the balls in this box at random

If a red ball is selected, what is the probability that this ball is from the first box?

Solution

E : a red ball is selected

$\overline{E}$: a blue ball is selected

F : a ball is selected from the first box

$\overline{F}$: a ball is selected from the second box

$$p(F \mid E) = \frac{p(F \cap E)}{p(E)}$$

$p(F \mid E)=p(F\cap E)/p(E) = 49/76 \approx 0.645$

$p(E \mid F)=p(E\cap F)/p(F)=7/9,\ p(F)=p(\overline{F})=1/2$

$\rightarrow p(E\cap F)=7/18$

$p(E \mid \overline{F})=p(E\cap \overline{F})/p(\overline{F})= 3/7 \rightarrow p(E\cap \overline{F})=3/14$

$E=(E\cap F)\ U\ (E\cap \overline{F}) \rightarrow p(E)=p(E\cap F)+p(E\cap \bar{F})=38/63$

Bayes Theorem

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

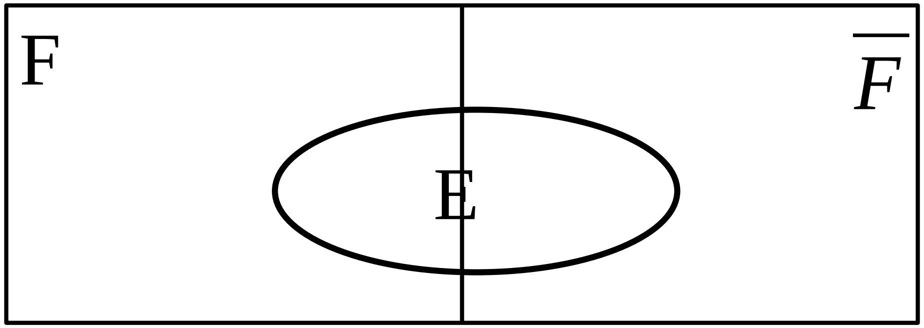
- $P(h)$ = prior probability of hypothesis h
- $P(D)$ = prior probability of training data D
- $P(h|D)$ = probability of h given D
- $P(D|h)$ = probability of D given h

Bayes Theorem

Suppose that E and F are events from a sample space S such that $p(E)\neq 0$ and $p(F)\neq 0$. Then

$$p(F \mid E) = \frac{p(E \mid F)\, p(F)}{p(E \mid F)\, p(F) + p(E \mid \overline{F})\, p(\overline{F})}$$

Note: F and $\overline{F}$ are mutually exclusive and cover the entire sample space S, that is $F \cup \overline{F} = S$.



Example

- Suppose one person in 100,000 has a particular rare disease for which there is a fairly accurate diagnostic test. This test is correct 99% of the time when given to someone with the disease; it is correct 99.5% of the time when given to someone who does not have the disease.

a) Only 0.2% of people who test positive actually have the disease.

b) 99.99999% of people who test negative really do not have the disease.

$$p(F \mid E) = \frac{p(E \mid F)p(F)}{p(E \mid F)p(F) + p(E \mid \overline{F})p(\overline{F})} = \frac{0.99 \bullet 0.00001}{0.99 \bullet 0.00001 + 0.005 \bullet 0.99999} \approx 0.002$$

$$p(\overline{F} \mid \overline{E}) = \frac{p(\overline{E} \mid \overline{F})p(\overline{F})}{p(\overline{E} \mid \overline{F})p(\overline{F}) + p(\overline{E} \mid F)p(F)} = \frac{0.995 \bullet 0.99999}{0.995 \bullet 0.99999 + 0.01 \bullet 0.00001} \approx 0.9999999$$

Choosing Hypotheses

- For each hypothesis h in H, calculate the posterior probability

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

- **Maximum A Posteriori** hypothesis h_{MAP}, Output the hypothesis h_{MAP} with the highest posterior probability

$$h_{MAP} =_{h \in H} P(h|D)$$

Choosing Hypotheses

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

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

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

Maximum Likelihood (ML)

- If assume $P(h_i)=P(h_j)$ for all h_i and h_j, then can further simplify, and choose the **Maximum likelihood** (ML) hypothesis

$$h_{ML} = \arg\max_{h_i \in H} P(D|h_i)$$

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, 0.008 of the entire population have this cancer

Example

$$P(cancer) = 0.008 \qquad P(\neg cancer) = 0.992$$

$$P(+|cancer) = 0.98 \qquad P(-|cancer) = 0.02$$

$$P(+|\neg cancer) = 0.03 \qquad P(-|\neg cancer) = 0.97$$

$$P(+|cancer) \cdot P(cancer) = 0.98 \cdot 0.008 = 0.0078$$

$$P(+|\neg cancer) \cdot P(\neg cancer) = 0.03 \cdot 0.992 = 0.0298$$

$$h_{MAP} = \neg cancer$$

Normalization

$$\frac{0.0078}{0.0078+0.0298}=0.20745 \qquad \frac{0.0298}{0.0078+0.0298}= 0.79255$$

- The result of Bayesian inference depends strongly on the prior probabilities, which must be available in order to apply the method

Naive Bayes Classifier

- A new instance is described by the tuple of attribute values $<a_1,\ldots,a_n>$.
- The learner is asked to predict the target value, or classification, for the new instance.
- The Bayesian approach to classifying the new instance is to assign the most probable target value, v_{MAP},

$$v_{MAP} = \underset{v_j}{\arg\max}\, P(v_j \mid a_1,\ldots,a_n)$$

- Bayes theorem to rewrite

$$\begin{aligned} v_{MAP} &= \underset{v_j \in V}{\arg\max} \frac{P(a_1,\ldots,a_n \mid v_j)P(v_j)}{P(a_1,\ldots,a_n)} \\ &= \underset{v_j \in V}{\arg\max}\, P(a_1,\ldots,a_n \mid v_j)P(v_j) \end{aligned}$$

Naive Bayes Classifier

- Naive Bayes Classifier

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i \mid v_j)$$

- Estimate $P(a_i|v_j)$ is much easier than estimate $P(a_1,\ldots,a_n|v_j)$

- Bayes assumption of conditional *independence* is satisfied, this Naive Bayes classification v_{NB} is identical to the MAP classification

Naive Bayes Classifier

Day	Outlook	Temperature	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	Cool	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

Naive Bayes Classifier

- Play tennis example provides 14 instances. Each instance has four attributes <sunny, cool, high, strong>

$$v_{NB} = \underset{v_j \in \{yes,no\}}{\arg\max} P(v_j) \prod_i P(a_i \mid v_j) = \underset{v_j \in \{yes,no\}}{\arg\max} P(v_j) P(sunny \mid v_j) P(cool \mid v_j) P(high \mid v_j) P(strong \mid v_j)$$

  - $P(yes)=9/14=0.64$
  - $P(no)=5/14=0.36$
  - $P(strong|yes)=3/9=0.33$
  - $P(strong|no)=3/5=0.60$
  - $P(yes)P(sunny|yes)P(cool|yes)P(high|yes)P(strong|yes)=0.0053$
  - $P(no)P(sunny|no)P(cool|no)P(high|no)P(strong|no)=0.0206$
  - $v_{NB}=no$

Example

- Example: Play Tennis

PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
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	Cool	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

Example

- Learning Phase

Outlook	Play=*Yes*	Play=*No*
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Temperature	Play=*Yes*	Play=*No*
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Humidity	Play=*Yes*	Play=No
High	3/9	4/5
Normal	6/9	1/5

Wind	Play=*Yes*	Play=*No*
Strong	3/9	3/5
Weak	6/9	2/5

$P(\text{Play}=Yes) = 9/14$ $\quad$ $P(\text{Play}=No) = 5/14$

Example

- Test Phase
  - Given a new instance,

  x'=(Outlook=*Sunny*, Temperature=*Cool*, Humidity=*High*, Wind=*Strong*)
  - Look up tables

P(Outlook=*Sunny*|Play=*Yes*) = 2/9	P(Outlook=*Sunny*|Play=*No*) = 3/5
P(Temperature=*Cool*|Play=*Yes*) = 3/9	P(Temperature=*Cool*|Play==*No*) = 1/5
P(Huminity=*High*|Play=*Yes*) = 3/9	P(Huminity=*High*|Play=*No*) = 4/5
P(Wind=*Strong*|Play=*Yes*) = 3/9	P(Wind=*Strong*|Play=*No*) = 3/5
P(Play=*Yes*) = 9/14	P(Play=*No*) = 5/14

  - Use the MAP rule to calculate Yes or No

P(*Yes*|**x**'): [P(*Sunny*|*Yes*)P(*Cool*|*Yes*)P(*High*|*Yes*)P(*Strong*|*Yes*)]P(Play=*Yes*) = 0.0053

P(*No*|**x**'): [P(*Sunny*|No) P(*Cool*|No)P(*High*|*No*)P(*Strong*|*No*)]P(Play=*No*) = 0.0206

Given the fact P(*Yes*|**x**') < P(*No*|**x**'), we label **x**' to be "*No*".