The Bayesian Learning

Rodrigo Fernandes de Mello

Invited Professor at Télécom ParisTech

Associate Professor at Universidade de São Paulo, ICMC, Brazil

http://www.icmc.usp.br/~mello

mello@icmc.usp.br

First step

- What is **Conditional Probability?**
  - The probability an event A occurs given event B has happened

$$P(A|B)$$

  - We represent it in form:

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

  - Or:

$$P(B \cap A) = P(A \mid B)P(B)$$

For example

- Conditional Probability:
  - When there is dependency:
    - In Bayesian Learning, we assume that one attribute depends on the values of another (or others)
  - For example:

Day	Outlook	Temperature	Moisture	Wind	Play Tennis
D1	Sunny	Warm	High	Weak	No
D2	Sunny	Warm	High	Strong	No
D3	Cloudy	Warm	High	Weak	Yes
D4	Rainy	Pleasant	High	Weak	Yes
D5	Rainy	Cold	Normal	Weak	Yes
D6	Rainy	Cold	Normal	Strong	No
D7	Cloudy	Cold	Normal	Strong	Yes
D8	Sunny	Pleasant	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rainy	Pleasant	Normal	Weak	Yes
D11	Sunny	Pleasant	Normal	Strong	Yes
D12	Cloudy	Pleasant	High	Strong	Yes
D13	Cloudy	Warm	Normal	Weak	Yes
D14	Rainy	Pleasant	High	Strong	No

For example

- Conditional Probability:
  - What is the probability that an event A occurs given B?

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

  - Assume B equals to Moisture = High

Day	Outlook	Temperature	Moisture	Wind	Play Tennis
D1	Sunny	Warm	High	Weak	No
D2	Sunny	Warm	High	Strong	No
D3	Cloudy	Warm	High	Weak	Yes
D4	Rainy	Pleasant	High	Weak	Yes
D5	Rainy	Cold	Normal	Weak	Yes
D6	Rainy	Cold	Normal	Strong	No
D7	Cloudy	Cold	Normal	Strong	Yes
D8	Sunny	Pleasant	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rainy	Pleasant	Normal	Weak	Yes
D11	Sunny	Pleasant	Normal	Strong	Yes
D12	Cloudy	Pleasant	High	Strong	Yes
D13	Cloudy	Warm	Normal	Weak	Yes
D14	Rainy	Pleasant	High	Strong	No

For example

- Conditional Probability:
  - Two possible values can be computed for A:
    - Play Tennis = Yes
    - Play Tennis = No
  - Therefore we have:

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

$$P(Jogar\ Tênis = Sim | Umidade = Alta) = \frac{P(Jogar\ Tênis = Sim \cap Umidade = Alta)}{P(Umidade = Alta)}$$

$$P(Jogar\ Tênis = Não | Umidade = Alta) = \frac{P(Jogar\ Tênis = Não \cap Umidade = Alta)}{P(Umidade = Alta)}$$

For example

- Conditional Probability:
  - What are the probabilities?

$$P(Umidade = Alta) = \frac{7}{14} = 0.5$$

$$P(Jogar\ Tênis = Sim \cap Umidade = Alta) = \frac{3}{14} = 0.214$$

$$P(Jogar\ Tênis = Não \cap Umidade = Alta) = \frac{4}{14} = 0.286$$

  - Thus:

$$P(Jogar\ Tênis = Sim | Umidade = Alta) = \frac{\frac{3}{14}}{\frac{7}{14}} = 0.428$$

$$P(Jogar\ Tênis = Não | Umidade = Alta) = \frac{\frac{4}{14}}{\frac{7}{14}} = 0.571$$

For example

- Conditional Probability:
  - Conclusion:
    - Knowing the moisture is high, we can infer the probabilities:
      - Play Tennis = Yes is equal to 42.8%
      - Play Tennis = No is equal to 57.1%

The Bayes Theorem

- Bayes Theorem:

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

Assim: $P(B \cap A) = P(A|B) \cdot P(B)$

Como $P(B \cap A) = P(A \cap B)$ logo:

$$P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$

e chegamos ao Teorema de Bayes: $P(A|B) = \dfrac{P(B|A) \cdot P(A)}{P(B)}$

Bayes Theorem

- In machine learning, we wish:
  - The best hipothesis h_{MAP} contained in space H given we have an observable training set D
    - Thus proceeding with the substitution in:

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

    - We have (for a hipothesis h in H):

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

Bayes Theorem

- However, to obtain h_{MAP} we need to compute:

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

- We refer to h_{MAP} as the hipothesis with the Maximum A Posteriori Probability (MAP)
  - i.e., the hypothesis that produces best results to unseen examples given the training on set D

What most people use in practice: The Naive Bayes Classifier

Naive Bayes Classifier

- According to the Bayes Theorem:
  - We attempt to classify an unseen example according to the most probable class, given a set of attributes <a_1, a_2, ..., a_n>:

$$v_{MAP} = \arg\max_{v_j \in V} \frac{P(a_1, a_2, \ldots, a_n | v_j) P(v_j)}{P(a_1, a_2, \ldots, a_n)}$$

- Using the training set, we need to estimate:
  - Probability $P(v_j)$, which is simple to be estimated
  - However, assuming the training set has a limited size:
    - It becomes difficult to estimate $P(a_1, a_2, \ldots, a_n | v_j)$, because there is possibly few or no identical occurrence in the training set (due to its size, i.e., numbers of examples)
    - This second probability could be estimated if and only if the training set were huge!

Naive Bayes Classifier

- The Naive Bayes Classifier simplifies this process:
  - It assumes that attributes are independent on each other
  - In other words, the probability of observing $a_1, a_2, \ldots, a_n$ is given by the product of the individual probabilities of attributes:

$$P(a_1, a_2, \ldots, a_n | v_j) = \prod_i P(a_i | v_j)$$

  - Thus, Naive Bayes simplifies the classification process as follows:

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

    - Using Naive Bayes, we observe there is no explicit search for a hypothesis
      - The hypothesis is always formulated by counting frequencies according to the query example (unseen example)

Naive Bayes Classifier: Example

- Consider the training set:

Day	**Outlook**	**Temperature**	**Moisture**	**Wind**	**Play Tennis**
D1	Sunny	Warm	High	Weak	No
D2	Sunny	Warm	High	Strong	No
D3	Cloudy	Warm	High	Weak	Yes
D4	Rainy	Pleasant	High	Weak	Yes
D5	Rainy	Cold	Normal	Weak	Yes
D6	Rainy	Cold	Normal	Strong	No
D7	Cloudy	Cold	Normal	Strong	Yes
D8	Sunny	Pleasant	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rainy	Pleasant	Normal	Weak	Yes
D11	Sunny	Pleasant	Normal	Strong	Yes
D12	Cloudy	Pleasant	High	Strong	Yes
D13	Cloudy	Warm	Normal	Weak	Yes
D14	Rainy	Pleasant	High	Strong	No

- Suppose the unseen example:

<Outlook=Sunny, Temperature=Cold, Moisture=High, Wind=Strong>

- Our task is to predict Yes or No to the concept “Play Tennis”

Naive Bayes Classifier: Example

- In this case we have:

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

- Thus:

$$v_{NB} = \arg\max_{v_j \in V} P(v_j)$$
$$P(Panorama = Ensolarado|v_j)P(Temperatura = Fria|v_j)$$
$$P(Umidade = Alta|v_j)P(Vento = Forte|v_j)$$

- Computing:

$$P(Vento = Forte|Jogar\ Tênis = Sim) = 3/9$$
$$P(Vento = Forte|Jogar\ Tênis = Não) = 3/5$$
$$P(Umidade = Alta|Jogar\ Tênis = Sim) = 3/9$$
$$P(Umidade = Alta|Jogar\ Tênis = Não) = 4/5$$
$$P(Temperatura = Fria|Jogar\ Tênis = Sim) = 3/9$$
$$P(Temperatura = Fria|Jogar\ Tênis = Não) = 1/5$$
$$P(Panorama = Ensolarado|Jogar\ Tênis = Sim) = 2/9$$
$$P(Panorama = Ensolarado|Jogar\ Tênis = Não) = 3/5$$

Naive Bayes Classifier: Example

- In which:

$$P(v_j = Sim) = 9/14$$
$$P(v_j = Não) = 5/14$$

- Thus:

$$P(Sim)P(Ensolarado|Sim)P(Fria|Sim)P(Alta|Sim)P(Forte|Sim) = 0.0053$$
$$P(Não)P(Ensolarado|Não)P(Fria|Não)P(Alta|Não)P(Forte|Não) = 0.0206$$

- Normalizing that, we have the probability for “Play Tennis”=No is 0.795, i.e., there is a 79.5% chance there will be no game
  - Observe Naive Bayes works on discrete data!!!

Naive Bayes Classifier

- Let's implement Naive Bayes...

Naive Bayes Classifier

- Commonly used to classify documents
  - 20 Newsgroups, Reuters
- Questions?