Naive Bayes Classifier

CSCI 4352 Machine Learning

Dongchul Kim, Ph.D.

Department of Computer Science

Bayesian Methods

- Our focus this lecture
  - Learning and classification methods based on probability theory.
  - Bayes theorem plays a critical role in probabilistic learning and classification.
  - Uses *prior* probability of each category given no information about an item.
  - Categorization produces a *posterior* probability distribution over the possible categories given a description of an item.

Basic Probability Formulas

- Product rule $$P(A \wedge B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$
- Sum rule $$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$
- Bayes theorem
$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$
- Theorem of total probability, if event A_i is mutually exclusive and probability sum to 1

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

Bayes Theorem

- Given a hypothesis *h* and data *D* which bears on the hypothesis:

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

- *P(h)*: independent probability of *h*: *prior probability*
- *P(D)*: independent probability of *D*
- *P(D|h)*: conditional probability of *D* given h: *likelihood*
- *P(h|D)*: conditional probability of *h* given *D*: *posterior probability*

Maximum A Posterior

- Based on Bayes Theorem, we can compute the *Maximum A Posterior* (MAP) hypothesis for the data
- We are interested in the best hypothesis for some space H given observed training data D.

$$
\begin{aligned}
h_{MAP} &\equiv \underset{h \in H}{\operatorname{argmax}} P(h \mid D) \\
&= \underset{h \in H}{\operatorname{argmax}} \frac{P(D \mid h) P(h)}{P(D)} \\
&= \underset{h \in H}{\operatorname{argmax}} P(D \mid h) P(h)
\end{aligned}
$$

H: set of all hypothesis.

Note that we can drop *P(D)* as the probability of the data is constant (and independent of the hypothesis).

Maximum Likelihood

- Now assume that all hypotheses are equally probable a priori, i.e., $P(h_i) = P(h_j)$ for all h_i, h_j belong to H.
- This is called assuming a *uniform prior*. It simplifies computing the posterior:

- This hypothesis is called the *maximum likelihood hypothesis.*

$$h_{ML} = \arg\max_{h \in H} P(D \mid h)$$

Desirable Properties of Bayes Classifier

- *Incrementality:* with each training example, the prior and the likelihood can be updated dynamically: flexible and robust to errors.
- *Combines prior knowledge and observed data:* prior probability of a hypothesis multiplied with probability of the hypothesis given the training data
- *Probabilistic hypothesis:* outputs not only a classification, but a probability distribution over all classes

Bayes Classifiers

Assumption: training set consists of instances of different classes described c_j as conjunctions of attributes values

Task: Classify a new instance *d* based on a tuple of attribute values into one of the classes $c_j \in C$

Key idea: assign the most probable class c_{MAP} using Bayes Theorem.

$$c_{MAP} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j \mid x_1, x_2, \ldots, x_n)$$

$$= \underset{c_j \in C}{\operatorname{argmax}} \frac{P(x_1, x_2, \ldots, x_n \mid c_j) P(c_j)}{P(x_1, x_2, \ldots, x_n)}$$

$$= \underset{c_j \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c_j) P(c_j)$$

Parameters estimation

- $P(c_j)$
  - Can be estimated from the frequency of classes in the training examples.
- $P(x_1, x_2, \ldots, x_n | c_j)$
  - O(|X|^n•|C|) parameters
  - Could only be estimated if a very, very large number of training examples was available.
- Independence Assumption: attribute values are conditionally independent given the target value: ***naïve** Bayes*.

$$P(x_1, x_2, \ldots, x_n \mid c_j) = \prod_i P(x_i \mid c_j)$$

$$c_{NB} = \operatorname*{arg\,max}_{c_j \in C} P(c_j) \prod_i P(x_i \mid c_j)$$

Properties

- Estimating $P(x_i \mid c_j)$ instead of $P(x_1, x_2, \ldots, x_n \mid c_j)$ greatly reduces the number of parameters (and the data sparseness).
- The learning step in Naïve Bayes consists of

estimating $P(x_i \mid c_j)$ and $P(c_j)$ based on the frequencies in the training data

- An unseen instance is classified by computing the class that maximizes the posterior
- When conditioned independence is satisfied, Naïve Bayes corresponds to MAP classification.

Example. ‘Play Tennis’ data

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	*No*
Day2	Sunny	Hot	High	Strong	*No*
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	*No*
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	*No*
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	*No*

Question: For the day <sunny, cool, high, strong>, what’s the play prediction?

Naive Bayes solution

Classify any new datum instance $\mathbf{x}=(a_1, \ldots a_T)$ *as:*

$$h_{Naive\ Bayes} = \arg\max_h P(h)P(\mathbf{x} \mid h) = \arg\max_h P(h)\prod_t P(a_t \mid h)$$

- To do this based on training examples, we need to estimate the parameters from the training examples:
  - For each target value (hypothesis) *h*

$$\hat{P}(h) := \text{estimate } P(h)$$

  - For each attribute value a_t of each datum instance

$$\hat{P}(a_t \mid h) := \text{estimate } P(a_t \mid h)$$

Based on the examples in the table, classify the following datum ***x***:

x=(Outl=Sunny, Temp=Cool, Hum=High, Wind=strong)

- That means: Play tennis or not?

$$h_{NB} = \underset{h \in [yes,no]}{\arg\max}\, P(h)P(\mathbf{x} \mid h) = \underset{h \in [yes,no]}{\arg\max}\, P(h)\prod_t P(a_t \mid h)$$

$$= \underset{h \in [yes,no]}{\arg\max}\, P(h)P(Outlook = sunny \mid h)P(Temp = cool \mid h)P(Humidity = high \mid h)P(Wind = strong \mid h)$$

- Working:

$$P(PlayTennis = yes) = 9/14 = 0.64$$
$$P(PlayTennis = no) = 5/14 = 0.36$$
$$P(Wind = strong \mid PlayTennis = yes) = 3/9 = 0.33$$
$$P(Wind = strong \mid PlayTennis = no) = 3/5 = 0.60$$
$$etc.$$
$$P(yes)P(sunny \mid yes)P(cool \mid yes)P(high \mid yes)P(strong \mid yes) = 0.0053$$
$$P(no)P(sunny \mid no)P(cool \mid no)P(high \mid no)P(strong \mid no) = \mathbf{0.0206}$$
$$\Rightarrow answer : PlayTennis(x) = no$$

~~Lab 23~~

~~Based on the examples in the table, classify the following datum ***x***:~~

~~x=(Outl=Rain, Temp=Mild, Hum=Normal, Wind=Weak)~~

~~You have to show how you calculated posterior, likelihood, and prior probability of naive bayes classifier. Submit a pdf file.~~

Underflow Prevention

- Multiplying lots of probabilities, which are between 0 and 1 by definition, can result in floating-point underflow.
- Since *log(xy) = log(x) + log(y),* it is better to perform all computations by summing logs of probabilities rather than multiplying probabilities.
- Class with highest final un-normalized log probability score is still the most probable.

$$c_{NB} = \underset{c_{\mathrm{j}} \in C}{\operatorname{argmax}} \log P(c_j) + \sum_{i \in positions} \log P(x_i \mid c_j)$$