Naïve Bayes Classifiers

Simple (naïve) classification methods based on Bayes rules
What does Naïve mean?

- “Naïve” refers to the (naïve) assumption that data attributes are independent
- The Bayesian method can still be optimal even when this attribute independency is violated (Domingos, P., and M. Pazzani. 1997)
Properties of Bayes Classifier

- *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.

$$c_{MAP} = \arg\max_{c_j \in C} P(c_j \mid x_1, x_2, \ldots, x_n)$$

$$= \arg\max_{c_j \in C} \frac{P(x_1, x_2, \ldots, x_n \mid c_j)P(c_j)}{P(x_1, x_2, \ldots, x_n)}$$

$$= \arg\max_{c_j \in C} P(x_1, x_2, \ldots, x_n \mid c_j)P(c_j)$$
Parameters estimation

- Use the frequencies in the data (MLE)
  - \( P(c_j) \)
    - Can be estimated from the frequency of classes in the training examples.
  - \( P(x_1, x_2, ..., x_n | c_j) \)
    - \( O(|X|^n \cdot |C|) \) parameters
    - Require large number of training examples

- Independence assumption: attribute values are conditionally independent given the target value: naïve Bayes.
  \[
P(x_1, x_2, ..., x_n | c_j) = \prod_i P(x_i | c_j)
\]
  \[
c_{NB} = \arg \max_{c_j \in C} \prod_i P(x_i | c_j)
\]
greatly reduces the number of parameters & data sparseness
Bayes classification

- An unseen instance is classified by computing the class that maximizes the posterior

\[
P(x_1, x_2, \ldots, x_n \mid c_j) \quad P(x_i \mid c_j)
\]

- When conditioned independence is satisfied, Naïve Bayes corresponds to MAP classification.

\[
P(c_j) \quad P(x_i \mid c_j)
\]
### Example: ‘play tennis’ data

<table>
<thead>
<tr>
<th>Day</th>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Wind</th>
<th>Play Tennis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day1</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>Day2</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>Day3</td>
<td>Overcast</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>Day4</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>Day5</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>Day6</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>Day7</td>
<td>Overcast</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>Day8</td>
<td>Sunny</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>Day9</td>
<td>Sunny</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>Day10</td>
<td>Rain</td>
<td>Mild</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>Day11</td>
<td>Sunny</td>
<td>Mild</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>Day12</td>
<td>Overcast</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>Day13</td>
<td>Overcast</td>
<td>Hot</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>Day14</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
</tbody>
</table>

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

Classify any new datum instance \( \mathbf{x} = (x_1, \ldots, x_n) \) as:

\[
C_{NB} = \arg \max_{c_j \in C} P(c_j) P(x_1, x_2, \ldots, x_n | c_j)
\]

\[
= \arg \max_{c_j \in C} P(c_j) \prod_{i}^{n} P(x_i | c_j)
\]

To do this based on training examples, we need to estimate the parameters from the training examples: \( P(c_j) \) & \( P(x_i | c_j) \)
Based on the examples in the table, classify the following datum $x$:
$x=(O=Sunny, T=Cool, H=High, W=strong)$

- That means: Play tennis or not?

$$c_{NB} = \arg \max_{c_j} P(c_j) \prod_i P(x_i|c_j)$$

$$= \arg \max_{c_j} P(c_j) P(O=sunny|c_j) P(T=cool|c_j) P(H=high|c_j) P(W=strong|c_j)$$

- 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.

$$P(yes)P(sunny \mid yes)P(cool \mid yes)P(high \mid yes)P(strong \mid yes) = 0.0053$$
$$P(no)P(sunny \mid no)P(cool \mid no)P(high \mid no)P(strong \mid no) = 0.0206$$

$\Rightarrow$ answer : $PlayTennis(x) = no$
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.

\[ c_{NB} = \arg\max_{c_j \in C} \log P(c_j) + \sum_{i \in \text{positions}} \log P(x_i \mid c_j) \]
Naïve Bayes for document classification

Relies on a very simple representation of document: Bag of words

\[ Y(\ ) = C \]

I like this movie. It is sweet, and makes me laugh. I will definitely recommend it. I’d like to watch it one more time!

Use all the words; or a subset of the words

C: like-it
dislike-it
Naïve Bayes for document classification

Relies on a very simple representation of document: Bag of words

\[ y(\text{like}, \text{laugh}, \text{recommend}, \ldots) = C \]
Naïve Bayes for interaction site prediction

Figure 7. Naïve Bayes Classifier with Model Parameters in the Form of CPTs

\[
\begin{align*}
\theta_C &= p(C|I) \\
\text{yes} & \quad p_2 & \quad 1 - p_2 \\
\text{no} & \quad p_3 & \quad 1 - p_3 \\
\theta_I &= p(I) \\
\text{yes} & \quad p_1 & \quad 1 - p_1 \\
\theta_H &= p(H|I) \\
\text{yes} & \quad p_4 & \quad 1 - p_4 \\
\text{no} & \quad p_5 & \quad 1 - p_5
\end{align*}
\]

http://www.ploscompbio.org/article/info:doi/10.1371/journal.pcbi.0030129
Naïve Bayes for Sequence Classification

- RDP classifier – for taxonomic assignment of ribosomal sequences
- A sequence is represented as a bag of k-mers (words)
RDP classifier

- Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy
- The Classifier uses a feature space consisting of all possible 8-base subsequences (words).
  - Word sizes between 6 and 9 bases were tested in preliminary experiments: Sizes of 8 and 9 bases gave nearly identical results, while sizes of 6 and 7 bases were less accurate.
- The position of a word in a sequence is ignored. As with text-based Bayesian classifiers, only those words occurring in the query contribute to the score.
RDP classifier: training

- **Word-specific priors:** \( W = \{ w_1, w_2, \ldots, w_d \} \)
  - Given \( N \) sequences, let \( n(w_i) \) be the number of sequences containing subsequence \( w_i \), \( P_i = [n(w_i) + 0.5]/(N + 1) \) (the likelihood of observing word \( w_i \) in an rRNA sequence).

- **Genus-specific conditional probabilities.**
  - For genus \( G \) with \( M \) sequences, let \( m(w_i) \) be the number of these sequences containing word \( w_i \). The conditional probability that a member of \( G \) contains \( w_i \): \( P(w_i|G) = [m(w_i) + P_i]/(M + 1). \)
  - Ignoring the dependency between words in an individual sequence, the joint probability of observing from genus \( G \) a (partial) sequence, \( S \), containing a set of words, \( V = \{ v_1, v_2, \ldots, v_f \} \) (\( V \subseteq W \)), was estimated as \( P(S|G) = \prod P(v_i|G). \)
RDP classifier: prediction

• By Bayes' theorem, the probability that an unknown query sequence, $S$, is a member of genus $G$ is $P(G|S) = P(S|G) \times P(G)/P(S)$, where $P(G)$ is the prior probability of a sequence being a member of $G$ and $P(S)$ the overall probability of observing sequence $S$ (from any genus).
• Assuming all genera are equally probable (equal priors), the constant terms $P(G)$ and $P(S)$ can be ignored.
• The sequence will be classified as a member of the genus giving the highest probability score, but we ignore the actual numerical probability estimate.