Naive Bayes

\[ \pi \pm 10^{-5} \]
AlphaGo vs. Lee Sedol update

Game 1
AlphaGo wins

Game 2
AlphaGo wins

Game 3
AlphaGo wins

Game 4
Lee Sedol wins
Reading Quiz

What machine learning task is the Naive Bayes algorithm used for?

a) dimensionality reduction
b) clustering
c) classification
d) regression
Naive Bayes: the setting

- Supervised learning (we know the correct output for each test point).
- Classification (outputs are discrete: class labels).
- Inputs also need to be discrete.
  - We will be estimating their probability from data.
Bayesian approach to classification

Use training data to estimate a probability for each label given an input:

$$P(label \mid input)$$

If we must output a single label, we choose the one with the highest probability.

Note: we’re changing the perspective here; there’s no longer one right answer.
Estimating probabilities from data

Suppose we flip a coin 10 times and observe: 7, 3

What do we believe to be the true $P(\cdot)$?

Now suppose we flip it 1000 times and observe: 700, 300
Estimating probabilities from data

**Key idea:** combine empirical frequency and prior probability.

Empirical frequency: \[
\frac{\text{observations of outcome}}{\text{total observations}} = \frac{n_H}{n}
\]

Prior for a coin toss: \[P(H) = \frac{1}{2}\]

Add \(m\) “observations” of the prior to the data: \[P(H) = \frac{n_H + \frac{1}{2}m}{n + m}\]
Estimating label probabilities

We want to output \( P(label \mid input) \)

- Conditional on a particular input point what is the probability of each label?

Estimating this empirically requires many observations of every possible input.

In such a case, we aren’t really learning: there’s no generalization to new data.

We want to generalize from many training points to get estimate a probability at an unobserved test point.
Naive Bayes assumption

**Key idea:** pretend each dimension of the input is independent, conditional on the output label.

For input \( x = (x_1, x_2, x_3) \) we can estimate \( P(x \mid l) = P(x_1 \mid l) \cdot P(x_2 \mid l) \cdot P(x_3 \mid l) \)

For each possible value of \( x_i \) and \( l \), we can compute an empirical frequency.

For example: \( P(x_1 = 2 \mid l = +1) = \frac{\text{count}(x_1 = 2 \land l = +1)}{\text{count}(l = +1)} \)
Bayes rule

We can estimate $P(x \mid l)$ empirically, but we actually want $P(l \mid x)$

We can get it using Bayes rule:

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$
Bayes rule applied

To compute $P(l \mid x_1) = \frac{P(x_1 \mid l) \cdot P(l)}{P(x_1)}$ from our data,

we need to estimate two more quantities from data: $P(x_1)$ and $P(l)$.

This means doing additional empirical estimates across our data set for each possible value of each input dimension and the label.
Pre-processing for naive Bayes

- We need to estimate the probability of each value for each dimension
  - For example: \( P(x_1 = 5) \)
- We need to estimate the probability of each label
  - For example: \( P(l = +1) \)
- We need to estimate the probability of each value for each dimension conditional on each label
  - For example: \( P(x_1 = 5 \mid l = -1) \)

All of these are estimated empirically, with some prior (usually uniform).
Classifying with naive Bayes

Given a new input $x = (x_1, x_2, x_3)$

Compute $P(label = l \mid x)$ for each possible label.

The component parts are computed by Bayes rule

$$P(l \mid x_1) = \frac{P(x_1 \mid l) \cdot P(l)}{P(x_1)}$$
Exercise: compute the posterior distributions

- $P(\text{sky} = \text{sunny})$
- $P(\text{temp} = \text{warm})$
- $P(\text{label} = \text{no})$
- $P(\text{label} = \text{yes})$
- $P(\text{label} = \text{yes} \mid \text{humidity} = \text{high})$
- $P(\text{label} = \text{no} \mid \text{wind} = \text{strong})$

<table>
<thead>
<tr>
<th>Example</th>
<th>Sky</th>
<th>AirTemp</th>
<th>Humidity</th>
<th>Wind</th>
<th>Water</th>
<th>Forecast</th>
<th>EnjoySport</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sunny</td>
<td>Warm</td>
<td>Normal</td>
<td>Strong</td>
<td>Warm</td>
<td>Same</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>Sunny</td>
<td>Warm</td>
<td>High</td>
<td>Strong</td>
<td>Warm</td>
<td>Same</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>Rainy</td>
<td>Cold</td>
<td>High</td>
<td>Strong</td>
<td>Warm</td>
<td>Change</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Sunny</td>
<td>Warm</td>
<td>High</td>
<td>Strong</td>
<td>Cool</td>
<td>Change</td>
<td>Yes</td>
</tr>
</tbody>
</table>