1 Logistic Regression

Create a logistic regression model that predicts low birthweight given the health and background of the mother. This is a built-in dataset called “birthwt”.

1. When you first import it into rattle, you will need to transform/recode the data: ignore “bwt” and make “race” categoric.
2. Keep the default split for test, train, and validation.
3. Learn a logistic linear classifier predicting “low” as the target.
4. What variables are important? What’s the relationship?
5. Product an error matrix on validation data.
6. Generate a csv with your predictions on the validation data (“score”).
7. Plot “age” vs. the regression.

```r
library(ggplot2)
ggplot(birthwt, aes(x=age, y=low)) + geom_point() +
  stat_smooth(method="glm", family="binomial", se=FALSE)
```
2 Naïve Bayes

Equation for decision function

\[ P(c|d) \propto P(c) \prod_{1 \leq i \leq n_d} P(w_i|c) \]

Our estimates for these priors and conditional probabilities:

\[ \hat{P}(c_j) = \frac{N_c + 1}{N + |C|} \]  \hspace{1cm} (1)

\[ \hat{P}(w|c) = \frac{T_{cw} + 1}{(\sum_{w' \in V} T_{cw'}) + |V|} \]  \hspace{1cm} (2)

Pretend that we saw these documents

<table>
<thead>
<tr>
<th>Color</th>
<th>Type</th>
<th>Origin</th>
<th>Stolen</th>
</tr>
</thead>
<tbody>
<tr>
<td>red</td>
<td>sports</td>
<td>domestic</td>
<td>Y</td>
</tr>
<tr>
<td>red</td>
<td>sports</td>
<td>domestic</td>
<td>N</td>
</tr>
<tr>
<td>red</td>
<td>sports</td>
<td>domestic</td>
<td>Y</td>
</tr>
<tr>
<td>yellow</td>
<td>sports</td>
<td>domestic</td>
<td>N</td>
</tr>
<tr>
<td>yellow</td>
<td>sports</td>
<td>imported</td>
<td>Y</td>
</tr>
<tr>
<td>yellow</td>
<td>suv</td>
<td>imported</td>
<td>N</td>
</tr>
<tr>
<td>yellow</td>
<td>suv</td>
<td>domestic</td>
<td>N</td>
</tr>
<tr>
<td>red</td>
<td>suv</td>
<td>imported</td>
<td>N</td>
</tr>
<tr>
<td>red</td>
<td>sports</td>
<td>imported</td>
<td>Y</td>
</tr>
</tbody>
</table>

Treat color, type, and origin as three independent random variables: \( f, t, \) and \( o \). The goal is to predict the class \( c \).

2.1 Estimation

Estimate the probability of

1. \( \hat{P}(f = \text{red} | c = \text{stolen}) \)
2. \( \hat{P}(f = \text{red} | c = \neg\text{stolen}) \)
3. \( \hat{P}(t = \text{suv} | c = \text{stolen}) \)
4. \( \hat{P}(t = \text{suv} | c = \neg\text{stolen}) \)
5. \( \hat{P}(o = \text{domestic} | c = \text{stolen}) \)
6. \( \hat{P}(o = \text{domestic} | c = \neg\text{stolen}) \)

2.2 Classification

Calculate the probability of a red domestic SUV being stolen.