1 General Questions

1. **What are the stages of the ML cycle? Which ones are iterative stages?**
   - Pre-processing
   - Feature-Extraction
   - Feature Selection
   - Machine Learning or Model creation
   - Evaluation and Model Selection

   All the stages are iterative, which means that you might come back to a previous stage in order to refine the process.

2. **What are the different types of learning?**
   - Supervised (Classification, Regression)
   - Unsupervised

3. **How would you describe the overfitting and underfitting phenomenon?**
   - Overfitting: Capturing noise in the data, because of a very complex model.
   - Underfitting: Unable to capture the patterns in the data, because of a very simple model.

2 Naive Bayes

2.1 Estimate probabities

1. $P(yes) = \frac{6}{10} = 0.6$
2. $P(red|yes) = \frac{3}{6} = 0.5$
3. $P(grandtourer|yes) = \frac{2}{6} = \frac{1}{3}$
4. $P(domestic|yes) = \frac{2}{6} = \frac{1}{3}$
5. $P(no) = \frac{4}{10} = 1 - P(yes) = 0.4$
6. $P(red|no) = \frac{1}{4} = 0.25$
7. $P(grandtourer|no) = \frac{2}{4} = 0.5$
8. $P(domestic|no) = \frac{3}{4} = 0.75$

2.2 Inference

We are asked to calculate $P(yes|red, gt, dom)$. Let's insert this in the formula given (equation 1) and ignore Z for the moment (which is why we repace "=" with "$\propto$").

$$\begin{aligned} P(yes|red, gt, dom) &\propto P(yes) \cdot P(red|yes) \cdot P(gt|yes) \cdot P(dom|yes) && (1)\\ &\propto 0.6 \cdot 0.5 \cdot \frac{1}{3} \cdot \frac{1}{3} && (2)\\ &\propto \frac{1}{30} && (3) \end{aligned}$$

For the normalization Z, we now also use the formula given in the assignment (equation 2). Here we sum over all possible values of y given the same attributes X, i.e.

$$Z = \sum_{y \in \{yes, no\}} P(y) \cdot P(red|y) \cdot P(gt|y) \cdot P(dom|y) \qquad (4)$$

For $y = yes$ we have done that already above. Now we still need to do it for $y = no$.

$$\begin{aligned} P(no|red, gt, dom) &\propto P(no) \cdot P(red|no) \cdot P(gt|no) \cdot P(dom|no) && (5)\\ &\propto 0.4 \cdot 0.25 \cdot 0.5 \cdot 0.75 && (6)\\ &\propto \frac{3}{80} && (7) \end{aligned}$$

From here, we can already see that it is more likely that the car is not being stolen as $\frac{1}{30} < \frac{3}{80}$. Since we are asked to calculate the exact probability, we can now easily do so with the following formula:

$$\begin{aligned} P(yes|red, gt, dom) &= \frac{\frac{1}{30}}{\frac{1}{30} + \frac{3}{80}} && (8)\\ &= \frac{8}{17} \approx 0.47. && (9) \end{aligned}$$

So the probability that a car with the given attributes is stolen is approximately 0.47

2.3 Benefits and downsides

Benefits

1. Less parameters
2. It is fast

3. It is easy

Downsides

1. Conditional independence is a strong assumption that might not hold in practice.

2.4 Derivation of Naive Bayes

The Bayes theorem states as follows:

$$P(y|X) = \frac{P(y)P(X|y)}{P(X)} \tag{10}$$

The nominator is equivalent to $P(y, X)$ and set $P(X) = Z$. This gives us

$$P(y|X) = \frac{1}{Z}P(y, X) \tag{11}$$

Using the following identity (chain rule of probabilities):

$$\begin{aligned} P(y, X) =& P(y) \cdot P(x_1|y) \cdot P(x_2|y, x_1) \cdot \\ & P(x_3|y, x_1, x_2) \cdot ... \cdot \\ & P(x_M|y, x_1, x_2, ..., x_{M-1}) \end{aligned}$$

and the assumption of conditional independence

$$\begin{aligned} P(y, X) :=& P(y) \cdot P(x_1|y) \cdot P(x_2|y) \cdot P(x_3|y) \cdot ... \cdot P(x_M|y) \\ :=& P(y) \prod_{i=1}^{M} P(x_i|y) \end{aligned} \tag{12}$$

we can obtain the formula for Naive bayes by plugging Equation 12 into Equation 11:

$$P(y|X) = \frac{1}{Z}P(y) \prod_{i=1}^{n} P(x_i|y) \tag{13}$$

3 Ranking Losses

1. We can reformulate this ranking problem as pairwise classification problem, where we create an auxiliary target y^* that compares two values x_1^* and x_2^* and whose output is computed as: $y^* = 1$ if the input x_2^* has a higher score than x_1^*, otherwise $y^* = 0$. We also create auxiliary predictions for the ML models $\hat{y}^* = \sigma(\hat{y}(x_2^*) - \hat{y}(x_1^*))$, where $\sigma(x) = \frac{1}{1+e^{-x}}$.

Thus, we formulate the loss as a logistic loss using these new auxiliary targets:

$$\mathcal{L} = L(y^*, \hat{y}^*) = -y^* \cdot \log(\hat{y}^*) - (1 - y^*) \cdot \log(1 - \hat{y}^*) \tag{14}$$

x_1^*	x_2^*	y^*	$\hat{y}_1^*$	$\hat{y}_2^*$	$L(y^*, \hat{y}_1^*)$	$L(y^*, \hat{y}_2^*)$
x_1	x_2	1	$\sigma(3-1) = 0.88$	$\sigma(3-2) = 0.73$	0.05	0.14
x_2	x_3	1	$\sigma(2-3) = 0.26$	$\sigma(7-3) = 0.98$	0.58	0.01
x_1	x_3	1	$\sigma(2-1) = 0.73$	$\sigma(7-2) = 0.99$	0.14	0.01

We explain the first row in the table:

- $y(\mathrm{x}_2) = 2, y(\mathrm{x}_1) = 1$ (ground truth)
- If we set $x_1^* = \mathrm{x}_1, x_2^* = \mathrm{x}_2$ (arbitrary decision), then $y(x_2^*) > y(x_1^*)$, thus $y^* = 1$.
- $\hat{y}_1^*(\mathrm{x}_1, \mathrm{x}_2) = \sigma(\hat{y}_1(\mathrm{x}_2) - \hat{y}_1(\mathrm{x}_1)) = \sigma(3-1)$
- $\hat{y}^*(\mathrm{x}_1, \mathrm{x}_2) = \sigma(\hat{y}_2(\mathrm{x}_2) - \hat{y}_2(\mathrm{x}_1)) = \sigma(3-2)$

2. After the previous point, it is clear that the best one is the second model (total loss = 0.15). If we use the squared error (SE), the first model would look like the best (SE of model 1 = 2 vs SE of model 2 = 18), but it is not the case. Remember that we do not care about the real value but we care about the ranking of the predicted scores.