Naïve Bayes Classifier

Classifiers Determine Posterior Probabilities

- When we build a model for classification, the output probabilities depend on the observations inputs.
- Essentially, we determine: "*Given* attributes of this observation, the predicted probability of success is …"

$$P(success|attributes)$$

This is called a **posterior probability**.

We might also consider the *prior probabilities* that someone has those attributes or that someone is successful.

Bayesian Classifiers

- **Bayesian Classifiers** are based on Bayes' theorem.

- **Naïve Bayes Classifiers** assume that the effect of the inputs are *independent* of one another.

Example: Looking at the effect of *Car Size* and *Car Color* on whether or not an accident occurred.

$$\boldsymbol{P}(Accident \mid Size = Medium \,\&\, Color = Blue) = \boldsymbol{P}(Accident \mid Size = Medium) * \boldsymbol{P}(Accident \mid Color = Blue)$$

Bayesian Classifiers

<u>Example</u>: Looking at the effect of *Car Size* and *Car Color* on whether or not an accident occurred.

$$\boldsymbol{P}(Accident \mid Size = Medium \,\&\, Color = Blue) = \boldsymbol{P}(Accident \mid Size = Medium) * \boldsymbol{P}(Accident \mid Color = Blue)$$

Size	Color	Accident
Medium	Blue	Yes
Medium	Red	Yes
Large	Red	Yes
Large	Blue	No
Large	Blue	No
Small	Red	No

Inputs/Output

- Inputs (for basic implementation)
  - Categorical variables
  - Normally distributed numeric variables
  - Class Target

- Output
  - Probabilities that a point belongs to each class.

Bayes' Theorem

- Let $\boldsymbol{x} = \{x_1, x_2, \dots, x_p\}$ be a sample observation with components representing values made on a set of *p* attributes.
  - **x** = {"Medium", "Blue"} in example from previous slide.
- Let C be target class variable, taking levels $\{c_1, c_2, \dots, c_L\}$
  - c_1= "Accident" and c_2= "No Accident" in previous example (L=2)
- We want to predict the **posterior probability P($c_i|\boldsymbol{x}$)**
  - The probability that a given observation belongs to each class, given that we know its attributes.
  - Bayes' Theorem: $$P(c_i|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|c_i)P(c_i)}{P(\boldsymbol{x})}$$

Independence Assumption

$$P(c_i|\boldsymbol{x}) = \frac{\boldsymbol{P(x|c_i)}P(c_i)}{P(\boldsymbol{x})}$$

$$\boldsymbol{P(x|c_i) = \prod_{k=1}^{p} P(x_k|c_i)}$$

Size	Color	Accident
Medium	Blue	Yes
Medium	Red	Yes
Large	Red	Yes
Large	Blue	No
Large	Blue	No
Small	Red	No

$$P(Medium\ \&\ Blue\ |Yes) = \underbrace{P(Medium|Yes)}_{\frac{2}{3}} * \underbrace{P(Blue|Yes)}_{\frac{1}{3}} = \frac{2}{9}$$

Independence Assumption

$$P(c_i|\boldsymbol{x}) = \frac{\frac{2}{9}P(c_i)}{P(\boldsymbol{x})}$$

Size	Color	Accident
Medium	Blue	Yes
Medium	Red	Yes
Large	Red	Yes
Large	Blue	No
Large	Blue	No
Small	Red	No

$$P(\boldsymbol{x}) = \prod_{k=1}^{p} P(x_k)$$

$$P(Medium\ \&\ Blue) = \underbrace{P(Medium)}_{\frac{2}{6}}\underbrace{P(Blue)}_{\frac{3}{6}} = \frac{1}{6}$$

Independence Assumption

$$P(`Yes'|`Medium' \,\&\, `Blue') = \frac{\frac{2}{9}\,P(c_i)}{\frac{1}{6}}$$

$$P(`Yes') = \frac{1}{2}$$

Size	Color	Accident
Medium	Blue	Yes
Medium	Red	Yes
Large	Red	Yes
Large	Blue	No
Large	Blue	No
Small	Red	No

Final Result

$$P(`Yes'|`Medium' \,\&\, `Blue') = \frac{\frac{2}{9} * \frac{1}{2}}{\frac{1}{6}} = \frac{2}{3}$$

but…what happens when we look at $\mathrm{P(`No'} \,|\, `Medium' \,\&\, `Blue')$?

Independence Assumption

$$P(\text{`}No'|\text{`}Medium' \,\&\, \text{`}Blue') = \frac{\boldsymbol{P(x|c_i)}P(c_i)}{P(\boldsymbol{x})}$$

Size	Color	Accident
Medium	Blue	Yes
Medium	Red	Yes
Large	Red	Yes
Large	Blue	No
Large	Blue	No
Small	Red	No

$$P(Medium \,\&\, Blue\ |No) = \underbrace{P(Medium|No)}_{0} * \underbrace{P(Blue|No)}_{\frac{2}{3}} = 0$$

Independence Assumption

$$P(`No' | `Medium' \ \& `Blue') = 0$$

$$P(`Yes' | `Medium' \ \& `Blue') = \frac{2}{3}$$

Size	Color	Accident
Medium	Blue	Yes
Medium	Red	Yes
Large	Red	Yes
Large	Blue	No
Large	Blue	No
Small	Red	No

- Our independence assumption makes these probabilities *estimates* of the truth.
- The final probabilities will not necessarily sum to 1.
- We'll run into problems when certain attributes do not occur for certain levels of the outcome.

Inputs/Output

- Inputs (for basic implementation)
  - **Categorical variables** – Can determine probabilities based on cross-tabulation of each variable with target variable
  - **Normally distributed numeric variables** – Can determine probabilities based on values of the normal (Gaussian) distribution with mean μ and variance σ which would be estimated from the data.

$$g(x_i, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

- Output
  - Probabilities that a point belongs to each class.

Laplace Correction (Laplace Estimator)

- We use the following estimation based on the class independence assumption.

$$\boldsymbol{P(x|c_i) = \prod_{k=1}^{p} P(x_k|c_i)}$$

- What happens if there is a class, c_i, and an attribute value x_k such that none of the samples in c_i have that attribute value?

- $P(x_k|c_i) = 0$ which means necessarily that $P(\boldsymbol{x}|c_i) = 0$, even if the probabilities for all the other attributes are very large!

Laplace Correction (Laplace Estimator)

- Simplest trick is to add a very small number to each cell in every crosstabulation.
- For our situation…

Size	Color	Accident
Medium	Blue	Yes
Medium	Red	Yes
Large	Red	Yes
Large	Blue	No
Large	Blue	No
Small	Red	No

$$P(Medium\ \&\ Blue\ |No) = \underbrace{P(Medium|No)}_{0} * \underbrace{P(Blue|No)}_{\frac{2}{3}} = 0$$

	Yes	No
Small	0	2
Medium	2	0
Large	1	1

Laplace Correction (Laplace Estimator)

- Simplest trick is to add a very small number to each cell in every crosstabulation.
- For our situation…

Size	Color	Accident
Medium	Blue	Yes
Medium	Red	Yes
Large	Red	Yes
Large	Blue	No
Large	Blue	No
Small	Red	No

$$P(Medium\ \&\ Blue\ |No) = \underbrace{P(Medium|No)}_{0} * \underbrace{P(Blue|No)}_{\frac{2}{3}} = 0$$

	Yes	No
Small	0 + 0.01	2 + 0.01
Medium	2 + 0.01	0 + 0.01
Large	1 + 0.01	1 + 0.01

Laplace Correction (Laplace Estimator)

- Simplest trick is to add a very small number to each cell in every crosstabulation.
- For our situation…

Size	Color	Accident
Medium	Blue	Yes
Medium	Red	Yes
Large	Red	Yes
Large	Blue	No
Large	Blue	No
Small	Red	No

$$P(Medium \,\&\, Blue \mid No) = \underbrace{P(Medium \mid No)}_{\frac{0.01}{3.03}} * \underbrace{P(Blue \mid No)}_{\frac{2}{3.03}} = 0.0022$$

	Yes	No
Small	0.01	2.01
Medium	2.01	0.01
Large	1.01	1.01

Laplace Correction (Laplace Estimator)

- This correction is known as a smoothing parameter.
- In large datasets, it is most commonly set = 1.
- It is a `hyperparameter' than can be tuned via cross-validation.

Advantages of Naïve Bayes

- Intuitive/Simple to explain and implement
- Can produce very good predictions
- Especially powerful on categorical variables and **text**
- Relatively fast computation time
- Robust to noise and irrelevant attributes

Disadvantages of Naïve Bayes

- Assumption that variables are independent and equally important for prediction is often faulty. This could lead to poor performance.
- Most easily applied with categorical or normally distributed variables – most software will make this assumption behind the scenes, even if variables not normally distributed – Careful!
- Requires more storage than other models - your training set tables essentially become your model.
- The more variables (counting levels) you have, the larger dataset required to make reliable estimates of each conditional probability
- Lose the ability to exploit interactions between variables
- Estimated probabilities are less trustworthy than predicted classes.