Naïve Bayes Classifier

Classifiers Determine Posterior Probabilities

Models determine: "*Given* attributes of this observation, the predicted probability of success is …"

$$P(success \mid attributes)$$

This is called a **posterior probability**.

We might also consider the *prior probabilities* that someone has those attributes or that someone is successful
(Simply P(attributes) or P(success)).

Bayesian Classifiers

- Bayesian Classifiers are **based on Bayes' theorem**.
- *Naïve* Bayes Classifiers **assume that the effect of the inputs are independent of one another.**
- When A, B are **independent events:**
    - $\mathbf{P(A\&B)=P(A)\cdot P(B)}$
    - $\mathbf{P(A\&B|C) = P(A|C)\cdot P(B|C)}$

Ex: Assuming Independence

$$P(Small \,\&\, Red) = P(Small) \cdot P(Red)$$

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

Ex: Assuming Independence

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

$$P(Small \ \& \ Red) = P(Small) \cdot P(Red) = \frac{3}{10}\frac{5}{10} = \frac{3}{20}$$

$$P(Small \ \& \ Red \,|\, Yes) = P(Small \,|\, Yes) \cdot P(Red \,|\, Yes)$$

Bayes' Theorem

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

Sample Calculation

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

Use Bayes' theorem to compute the posterior probability that a **Medium Blue** car experiences an accident.

$$P(Yes \mid Medium \;\&\; Blue)$$

Sample Calculation: $P(Yes \mid Medium \ \& \ Blue)$

$$P(c_i \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid c_i) P(c_i)}{P(x)} \longrightarrow P(x \mid c_i) = \prod_{k=1}^{p} P(x_k \mid c_i)$$

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

$$\begin{aligned} P(x \mid c_i) &= P(Med \ \& \ Blue \mid Yes) \\ &= P(Medium \mid Yes) \cdot P(Blue \mid Yes) \\ &= \frac{3}{6} \cdot \frac{2}{6} \\ &= \frac{1}{6} \end{aligned}$$

Sample Calculation: $P(Yes \mid Medium \ \& \ Blue)$

$$P(c_i \mid \mathbf{x}) = \frac{\frac{1}{6} P(c_i)}{P(x)}$$

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

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

$$\begin{aligned} P(x) &= P(Med \ \& \ Blue) \\ &= P(Medium) \cdot P(Blue) \\ &= \frac{3}{10} \cdot \frac{5}{10} \\ &= \frac{3}{20} \end{aligned}$$

Sample Calculation: $P(Yes \mid Medium \,\&\, Blue)$

$$P(c_i \mid \mathbf{x}) = \frac{\frac{1}{6} P(c_i)}{\frac{3}{20}}$$

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

$$P(c_i) = P(Yes) = \frac{6}{10}$$

Sample Calculation: $P(Yes \mid Medium \ \& \ Blue)$

Final Result

$$P(Yes \mid Medium \ \& \ Blue) = \frac{\frac{1}{6}\frac{6}{10}}{\frac{3}{20}} = \frac{2}{3}$$

but...what happens when we look at $P(No \mid Medium \ \& \ Blue)$?

Sample Calculation: $P(No \mid Medium \ \& \ Blue)$

$$P(c_i \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid c_i)P(c_i)}{P(x)}$$

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

Sample Calculation: $P(No \mid Medium \,\&\, Blue)$

$$P(c_i \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid c_i) P(c_i)}{P(x)}$$

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

$$\begin{aligned} P(x \mid c_i) &= P(Med \,\&\, Blue \mid No) \\ &= P(Medium \mid No) \cdot P(Blue \mid No) \\ &= 0 \cdot \frac{3}{4} \\ &= 0 \end{aligned}$$

Sample Calculation: $P(No \mid Medium \,\&\, Blue)$

$$P(c_i \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid c_i)P(c_i)}{P(x)}$$

$$P(x \mid c_i) = P(Med \,\&\, Blue \mid No)$$
$$= P(Medium \mid No) \cdot P(Blue \mid No)$$
$$= 0 \cdot \frac{3}{4}$$
$$= 0$$

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

Run into problems when certain attributes do not occur for certain levels of the outcome => predicted probabilities become exactly zero regardless of other attributes

Predicted Probabilities = 0 😱

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

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

- 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!

Solution: Laplace Correction (Laplace Estimator)

Laplace Correction (Laplace Estimator)

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

Simplest trick is to add a very small number to each cell in every crosstabulation.

$$P(x \mid c_i) = P(Med \,\&\, Blue \mid No)$$
$$= P(Medium \mid No) \cdot P(Blue \mid No)$$
$$= 0 \cdot \frac{3}{4}$$

	Yes	No
Small	0	2
Medium	2	0
Large	1	1

Laplace Correction (Laplace Estimator)

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

Simplest trick is to add a very small number to each cell in every crosstabulation.

$$P(x \mid c_i) = P(Med \,\&\, Blue \mid No)$$
$$= P(Medium \mid No) \cdot P(Blue \mid No)$$
$$= 0 \cdot \frac{3}{4}$$

	Yes	No
Small	1+0.01	2+0.01
Medium	3+0.01	0+0.01
Large	2+0.01	2+0.01

Laplace Correction (Laplace Estimator)

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

Simplest trick is to add a very small number to each cell in every crosstabulation.

$$P(x \mid c_i) = P(Med \,\&\, Blue \mid No)$$
$$= P(Medium \mid No) \cdot P(Blue \mid No)$$
$$= \frac{0.01}{4.03} \cdot \frac{3}{4} = 0.00186$$

	Yes	No
Small	1.01	2.01
Medium	3.01	0.01
Large	2.01	2.01

Laplace Correction (Laplace Estimator)

- This correction is known as a smoothing parameter.
- In large datasets, it is most commonly set $= 1$.
- Hyperparameter! Can be tuned via cross-validation.

Creating Output Probabilities

$$P(No \mid Medium \ \& \ Blue) = 0.00186$$

$$P(Yes \mid Medium \ \& \ Blue) = \frac{2}{3}$$

The final probabilities will not likely sum to 1 so we force them to by dividing by their sum

$$P(No \mid Medium \ \& \ Blue) = \frac{0.00186}{0.00186 + \frac{2}{3}} = 0.00278$$

$$P(Yes \mid Medium \ \& \ Blue) = \frac{\frac{2}{3}}{0.00186 + \frac{2}{3}} = 0.99722$$

Inputs/Output

Inputs (for basic implementation)

- **Categorical variables** – Determine probabilities based on cross-tabulation of each variable with target variable
- **Normally distributed numeric variables** – 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.

Summary of Naïve Bayes

Advantages

- 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

Summary of Naïve Bayes

Disadvantages

- **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 assume normality** behind the scenes, even if variables not normally distributed – Careful!
- **Requires more storage than other models** - your training set tables essentially become your model (slightly less storage than kNN).
- **More variables => more problems.** The more variables (including levels of categoricals), the larger the 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.