

# DATA SCIENCE

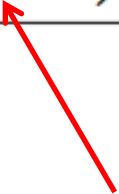
## NAIVE BAYES CLASSIFICATION

*Suppose we have a dataset with features  $x_1, \dots, x_n$  and a class label  $C$ .  
What can we say about classification using Bayes' theorem?*


$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

*Bayes' theorem can help us to determine the probability of a record belonging to a class, given the data we observe.*

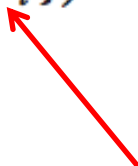
*This term is the **prior probability** of  $C$ . It represents the probability of a record belonging to class  $C$  before the data is taken into account.*

$$P(\text{class } C | \{x_i\}) = \frac{P(\{x_i\} | \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$


*This term is the **likelihood function**. It represents the joint probability of observing features  $\{x_i\}$  given that that record belongs to class  $C$ .*

$$P(\text{class } C | \{x_i\}) = \frac{P(\{x_i\} | \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$


*This term is the **normalization constant**. It doesn't depend on  $C$ , and is generally ignored.*

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$


*This term is the **posterior probability** of  $C$ . It represents the probability of a record belonging to class  $C$  after the data is taken into account.*

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

*The idea of Bayesian inference, then, is to **update** our beliefs about the distribution of  $C$  using the data (“evidence”) at our disposal.*

*Q: What piece of the puzzle we've seen so far looks like it could intractably difficult in practice?*

*A: Estimating the full likelihood function.*

$$P(\{x_i\}|C) = P(\{x_1, x_2, \dots, x_n\}|C)$$

*Observing this exactly would require us to have enough data for every possible combination of features to make a reasonable estimate.*

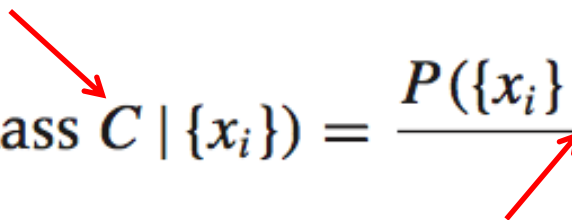
*Q: So what can we do about it?*

*A: Make a simplifying assumption. In particular, we assume that the features  $x_i$  are conditionally independent from each other:*

$$P(\{x_i\}|C) = P(\{x_1, x_2, \dots, x_n\}|C) \approx P(x_1|C) * P(x_2|C) * \dots * P(x_n|C)$$

*This “naïve” assumption simplifies the likelihood function to make it tractable.*




$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

*In summary, the **training phase** of the model involves computing the **likelihood function**, which is the conditional probability of each feature given each class.*

*The **prediction phase** of the model involves computing the **posterior probability** of each class given the observed features, and choosing the class with the highest probability.*