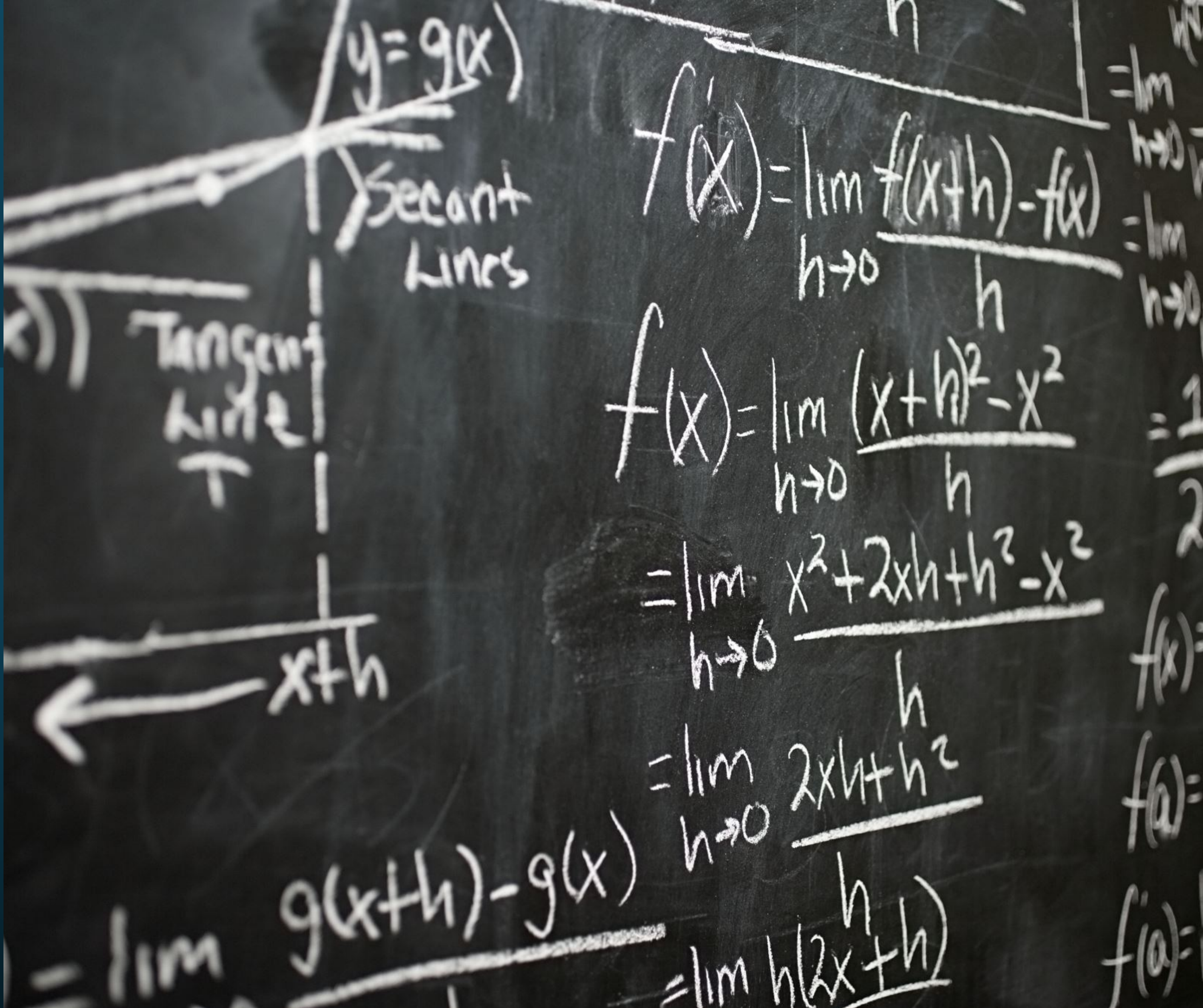
Based on Baye's theorem

Naïve Bayesian Classifier



Bayes' Theorem

- $$P(H|X) = \frac{P(X|H) \cdot P(H)}{P(X)}$$
- X: Data tuple
- H: Hypothesis that the data tuple X belongs to a specified class C
- $P(H|X)$: Posterior probability
- $P(H)$: Prior probability
- $P(X|H)$: Likelihood
- $P(X)$: Evidence

Naïve Bayesian Classification

- **Assumption:** Attributes are conditionally independent given the class (i.e. no dependence relationships among the attributes)
- **Formula:**

$$P(C_i|X) = \frac{P(X|C_i) \cdot P(C_i)}{P(X)}$$

- Since $P(X)$ is constant, we maximize:

$$P(X|C_i) \cdot P(C_i)$$

Naïve Bayes Classifier Formula

- $$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$
- x_k: value of attribute k
- n: number of attributes
- **Class label prediction:**

$$argmax_{C_i}\left[P(C_i)\prod_{k=1}^{n} P(x_k|C_i)\right]$$

Example: *AllElectronics* Customer Database

RID	age	income	student	credit_rating	buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Class Priors

- $C_1: buys\ computer = yes$
- $C_2: buys\ computer = no$
- The tuple we wish to classify is
  - $X = (age = youth, income = medium, student = yes, credit\ rating = fair)$
- $P(buys\ computer = yes) = \frac{9}{14} = 0.643$
- $P(buys\ computer = no) = \frac{5}{14} = 0.357$

Likelihoods

- $P(age = youth|buys\ computer = yes) = \frac{2}{9} = 0.22$
- $P(age = youth|buys\ computer = no) = \frac{3}{5} = 0.60$
- $P(income = medium|buys\ computer = yes) = \frac{4}{9} = 0.44$
- $P(income = medium|buys\ computer = no) = \frac{2}{5} = 0.40$
- $P(student = yes|buys\ computer = yes) = \frac{6}{9} = 0.667$
- $P(student = yes|buys\ computer = no) = \frac{1}{5} = 0.200$
- $P(credit\ rating = fair|buys\ computer = yes) = \frac{6}{9} = 0.667$
- $P(credit\ rating = fair|buys\ computer = no) = \frac{2}{5} = 0.400$

Likelihoods…

- $P(X|buys\ computer = yes) =$
  $P(age = youth|buys\ computer = yes) \times$
  $P(income = medium|buys\ computer = yes) \times$
  $P(student = yes|buys\ computer = yes) \times$
  $P(credit\ rating = fair|buys\ computer = yes)$
  $= 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.044$

- $P(X|buys\ computer = no) = P(age = youth|buys\ computer = no) \times$
  $P(income = medium|buys\ computer = no) \times$
  $P(student = yes|buys\ computer = no) \times$
  $P(credit\ rating = fair|buys\ computer = no)$
  $= 0.60 \times 0.40 \times 0.20 \times 0.40 = 0.019$

Classify Tuple

-

- $P(X|buys\ computer = yes) \cdot P(buys\ computer = yes) = 0.044 \times 0.643 = 0.028$
- $P(X|buys\ computer = no) \cdot P(buys\ computer = no) = 0.019 \times 0.357 = 0.007$
- **Prediction**: $buys\ computer = yes$ for tuple X.

Advantages and Disadvantages

- Fast, simple, works well with high-dimensional data
- Handles both continuous and discrete data
- Robust to noise
- Independence assumption often violated