Naïve Bayes Classifier

Instructor: Jing Wang

Department of ISOM

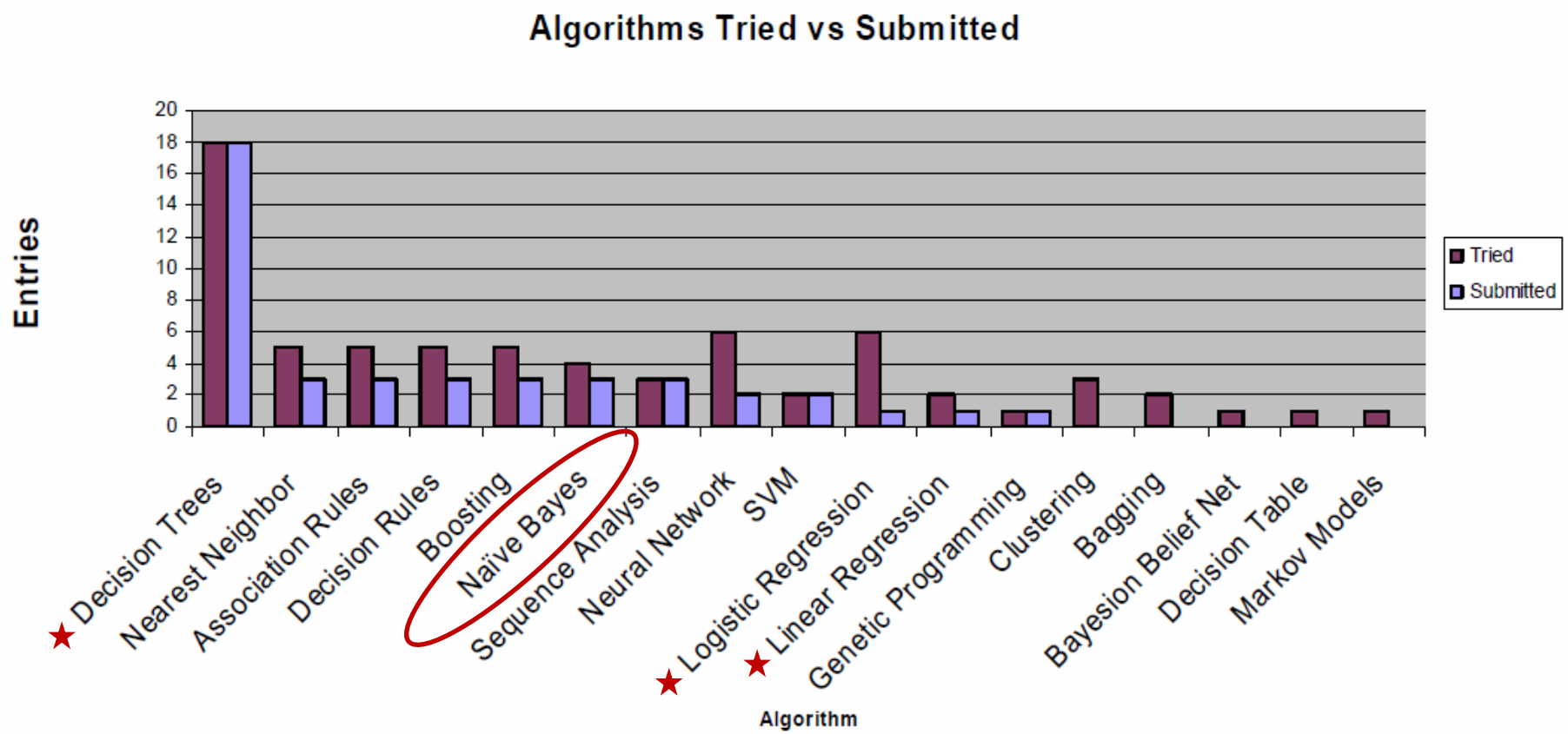
Spring 2023

HKUST
BUSINESS SCHOOL
香港科大商學院

Review

- Classification: predict a categorical outcome
- Classification Trees
  - Find informative attributes
  - Estimate probability (Y=1) using proportions
- Logistic regression
  - Linear model in log odds
  - Estimate probability (Y=1) in a more sophisticated way

Commonly Used Induction Algorithms



Naïve Bayes Classifier

- Find **$P(Y=1 \mid X_1, X_2, \ldots, X_m)$**
- Example
  - P(Y=Default |Income<30K, Education=College,.....)
  - P(Y=Spam email| X1="lottery", X2= "win", ...)
- Use probability theory to find the likelihood of the event of interest.

Probability and Conditional Probability

- Probability P(A): the probability of event A
  - e.g., suppose that in a certain city, 23 percent of the days are rainy. Thus, if you pick a random day, the probability that it rains on that day is 0.23: P(rain)=0.23
- Conditional probability P(A|B): the probability of event A, given event B
  - e.g., given that a day is cloudy, the chance that it rains increases to 62 percent. Then, the conditional probability of rain given cloudy is 0.62: P(rain|cloudy)=0.62

An Example

- The table below shows the results of a survey. How to calculate P(own a pet) and P(own a pet | female)?

Do you own a pet?

	Yes	No
female	8	6
male	5	7

- Of all the 26 respondents, 13 of them own a pet. So P(own a pet) = 13/26=0.50
- Of the 14 female respondents, 8 of them own a pet. So P(own a pet | female) = 8/14=0.57

Probability Chain Rule

- Chain rule: $P(A, B) = P(B|A)P(A) = P(A|B)P(B)$

	Own	Don't own
female	8	6
male	5	7

P(own a pet, female) = P(female) x P(own a pet | female)

= (14/26) x (8/14) = 8/26

P(own a pet, female) = P(own a pet) x P(female | own a pet)

= (13/26) x (8/13) = 8/26

Bayes' Theorem

- From chain rule, we have

$$P(A, B) = P(B|A)P(A) = P(A|B)P(B)$$

- Bayes' Theorem (or called Bayes' Rule)

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

Bayes' Theorem: Example 1

- Let Smart denotes smart students and GradeA denote students who get grade A. Assume that P(Smart)=0.3, and we believe P(GradeA|Smart) = 0.6, and now we learn that a student receives grade A. Can we estimate the probability that the student is smart, i.e., P(Smart|GradeA)=?

$$P(Smart|GradeA) = P(GradeA|Smart) \times P(Smart) / P(GradeA)$$

  - If P(GradeA) = 0.2, P(Smart|GradeA) = ? (0.6*0.3)/0.2 = 0.9
  - If P(GradeA) = 0.4, P(Smart|GradeA) = ? (0.6*0.3)/0.4 = 0.45

Bayes' Theorem: Example 2

- Suppose a disease pre-screening is 95% accurate. That is, if the patient has the disease, then the test will be positive with probability 0.95, i.e. P(positive|disease)=0.95. If the patient does not have the disease, then the test will be negative with probability 0.95, i.e. P(negative|no disease)=0.95.
- Suppose the probability of having this disease is 1 in 1000. P(disease)=0.001.
- Now suppose that a person gets a positive test result, what is the probability that he has the disease, P(disease|positive)=?

Use Bayes' Theorem in Data Mining

Likelihood of seeing evidence 'E' if 'H' is true

Prior probability of 'H'

$$P(H \mid E) = \frac{P(E \mid H) \times P(H)}{P(E)}$$

Posterior probability of 'H' given the evidence 'E'

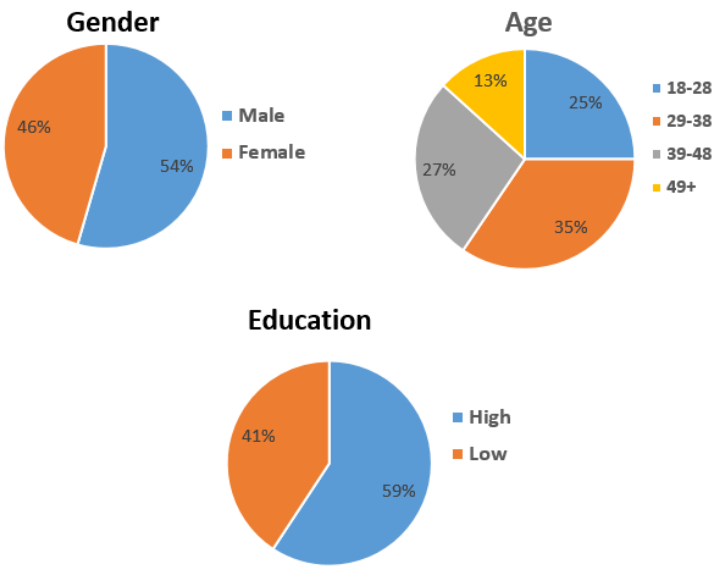
Prior probability that the evidence itself is true

E: Evidence (data)
H: Hypothesis (prediction)

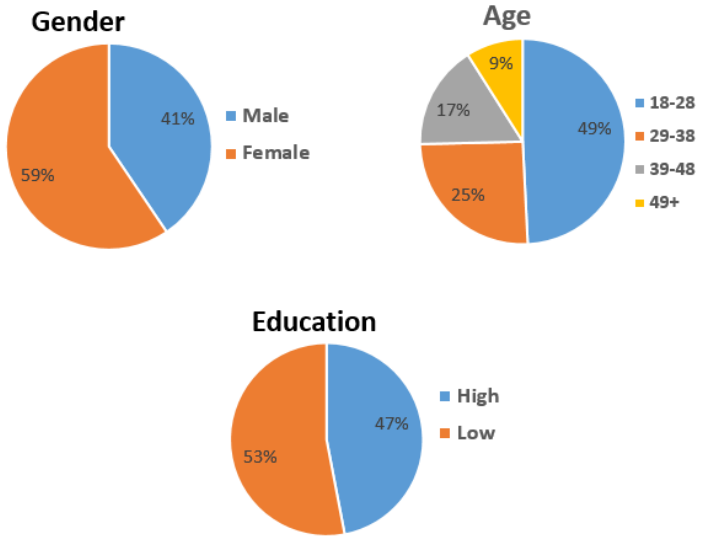
- The features (attribute values) observed are evidence of one hypothesis (class) or another.
  - Prior: existing knowledge
  - Posterior: updated knowledge given new observations

What a Churner/Non-Churner Looks Like?

Among **churning** customers, feature value distribution is shown as



Among **non-churning** customers, feature value distribution is shown as



- In modeling, we summarize feature patterns from training data for each class
- In use, when facing a new customer to be classified, we answer the question: ***"Which class is most likely to generate feature patterns exhibited by the customer?"***

Bayesian Classification

- Y : Target variable with binary values (1 or 0)
- $X_1, X_2, \ldots, X_k$: k attributes
- Objective: predict the value of Y given known $X_1, X_2, \ldots, X_k$
- Apply Bayes' rules

Hypothesis **(class)**

Evidence **(observed attributes)**

$$P(Y = 1 \mid X_1 = a_1, X_2 = a_2, \ldots X_k = a_k)$$

$$= \frac{P(X_1 = a_1, X_2 = a_2, \ldots X_k = a_k \mid Y = 1) \times P(Y = 1)}{P(X_1 = a_1, X_2 = a_2, \ldots X_k = a_k)}$$

Potential Challenge

$$P(X_1 = a_1, X_2 = a_2, \ldots X_k = a_k | Y = 1)$$

2^10 =1024

If we have 10 binary attributes, how many unique combinations of attribute values can we have? How about 20 binary attributes?

**As the number of attributes go large, it may not be possible to look for this probability from training data as you won't even have the same combination of attribute values for the test example as any training example.

Solution: assuming **conditional independence.**

Conditional Independence

- Event A="admitted to HKUST", event B="admitted to HKU", then A and B are not independent.
- Suppose event C: the student's entrance exam score is high, and both universities admit students only based on score. Then A is conditionally independent of B given C.
- P(HKUST | HKU, High Score) = P(HKUST | High Score)

Formally, we say two variables X1 is **conditionally independent** of X2 given Y, if and only if the probability distribution governing X1 is independent of the value of X2 given Y.

$$P(X_1 | X_2, Y) = P(X_1 | Y)$$

, 代表两个event同时发生

Other Examples

- Traffic and umbrellas are conditionally independent, given that it is raining.

- Height and vocabulary are conditionally independent, given age

Conditional Independence

- From $P(X_1 | X_2, Y) = P(X_1 | Y)$, we can also get

$$P(X_1, X_2 | Y) = P(X_1 | Y) \times P(X_2 | Y)$$

Can you figure it out?

$$P(X_1, X_2 | Y)$$

Chain rule

$$= P(X_1 | X_2, Y) \times P(X_2 | Y)$$

Conditional independence

$$= P(X_1 | Y) \times P(X_2 | Y)$$

Naïve Bayes Classifier

- The "naïve" assumption
  - The attribute values are conditionally independent, given the class.

$$P(X_1 = a_1, X_2 = a_2, \dots, X_k = a_k | Y = 1)$$
$$= P(X_1 = a_1 | Y = 1) \times \cdots \times P(X_k = a_k | Y = 1)$$

$$P(Y = 1 | X_1 = a_1, X_2 = a_2, \dots, X_k = a_k)$$
$$\propto P(X_1 = a_1, X_2 = a_2, \dots, X_k = a_k | Y = 1) \times P(Y = 1)$$
$$\propto P(X_1 = a_1 | Y = 1) \times \cdots \times P(X_k = a_k | Y = 1) \times P(Y = 1)$$

Calculating Probability of Evidence

- We ignore the denominator P(evidence) in previous few slides.
- How to calculate it? $P(X_1 = a_1, X_2 = a_2, \ldots X_k = a_k)$

$$
\begin{aligned}
&P(X_1 = a_1, X_2 = a_2, \ldots X_k = a_k) \\
&= P(X_1 = a_1, X_2 = a_2, \ldots X_k = a_k, Y = 1) \\
&\quad + P(X_1 = a_1, X_2 = a_2, \ldots X_k = a_k, Y = 0) \\
&= P(X_1 = a_1, X_2 = a_2, \ldots X_k = a_k | Y = 1) \times P(Y = 1) \\
&\quad + P(X_1 = a_1, X_2 = a_2, \ldots X_k = a_k | Y = 0) \times P(Y = 0) \\
&= P(X_1 = a_1 | Y = 1) \times \cdots \times P(X_k = a_k | Y = 1) \times P(Y = 1) \\
&\quad + P(X_1 = a_1 | Y = 0) \times \cdots \times P(X_k = a_k | Y = 0) \times P(Y = 0)
\end{aligned}
$$

Question

With the naïve (conditional independence) assumption, can you do the following?

$$P(X_1 = a_1, X_2 = a_2, \ldots X_k = a_k)$$

$$= P(X_1 = a_1) \times \cdots \times P(X_k = a_k)$$

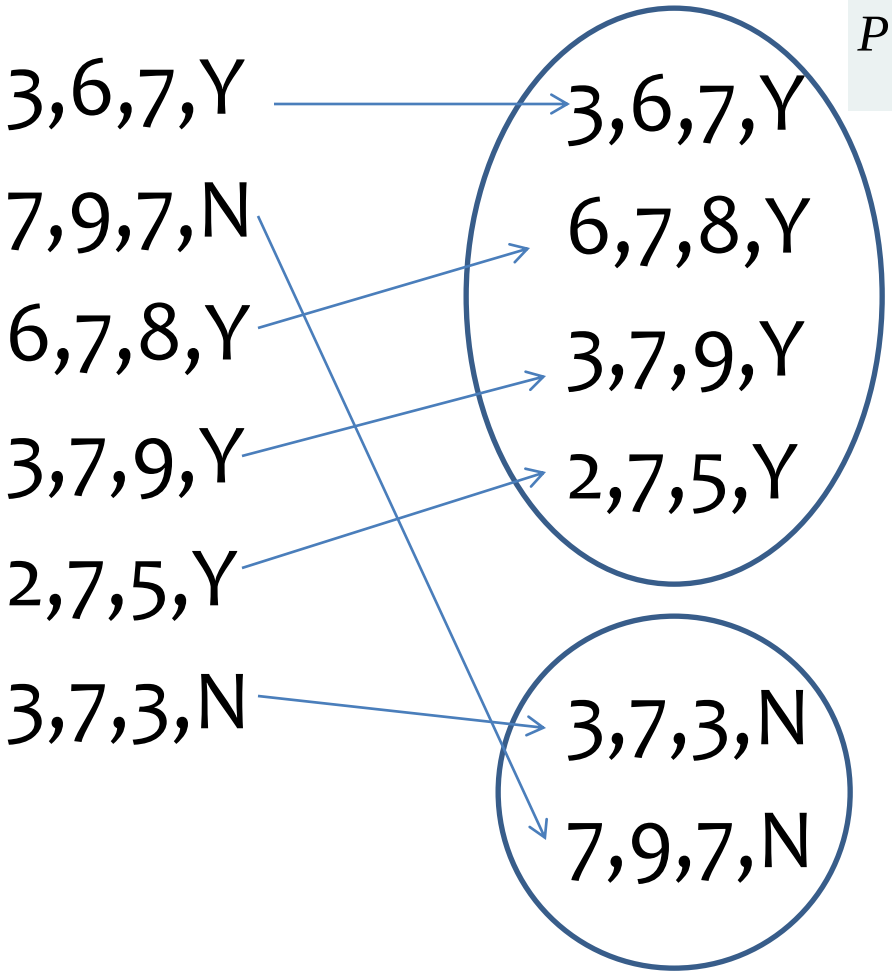
Calculating Probabilities

- Each element in the above product is estimated from training data (Data-driven strategies!!)

$$P(X_1 = a_1 | Y = 1) = \frac{count(X_1 = a_1 \; and \; Y = 1)}{count(Y = 1)}$$

$$P(X_1 = a_1 | Y = 0) = \frac{count(X_1 = a_1 \; and \; Y = 0)}{count(Y = 0)}$$

Exercise:



$$P(A_1 = a_1 \mid C = Y) = \frac{count(A_1 = a_1 \text{ and } C = Y)}{count(C = Y)}$$

Count (C = Y) = ?
Count (A_1=3 and C = Y) = ?
Count (A_1=6 and C = Y) = ?
Count (A_1=7 and C = Y) = ?

$P(A_1=a_1|C=Y)$ = ?

$P(A_1=3|c=Y)$ = ?
$P(A_1=6|c=Y)$ = ?
$P(A_1=7|c=Y)$ =?

Two-class problem where target variable **C** is either Y or N

Example

Outlook	Temp	Humidity	Windy	Class
sunny	hot	high	false	Don't Play
sunny	hot	high	true	Don't Play
overcast	hot	high	false	Play
rainy	mild	high	false	Play
rainy	cool	normal	false	Play
rainy	cool	normal	true	Don't Play
overcast	cool	normal	true	Play
sunny	mild	high	false	Don't Play
sunny	cool	normal	false	Play
rainy	mild	normal	false	Play
sunny	mild	normal	true	Play
overcast	mild	high	true	Play
overcast	hot	normal	false	Play
rainy	mild	high	true	Don't Play

Computing Conditional Probabilities

Outlook			Temperature			Humidity			Windy			Class	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2	3	hot	?	?	high	?	?	false	?	?	9	5
overcast	4	0	mild	?	?	normal	?	?	true	?	?		
rainy	3	2	cool	?	?								

	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2/9	3/5	hot	?	?	high	?	?	false	?	?	9/14	5/14
overcast	4/9	0/5	mild	?	?	normal	?	?	true	?	?		
rainy	3/9	2/5	cool	?	?								

P(Outlook = sunny |Class = no)

P(Class = yes)

Prediction Problem

- What class does this new example belong to? Class=yes or Class=no?

Outlook	Temp	Humidity	Windy	Class
sunny	cool	high	true	?

P(Class = yes | Outlook = sunny & Temp = cool & Humidity = high & Windy = true)=?

P(Class = no | Outlook = sunny & Temp = cool & Humidity = high & Windy = true)=?

Apply Bayes Rule

- Compute probability of Class=yes:

P(Class = yes | Outlook = sunny & Temp = cool & Humidity = high & Windy = true)
= P(Outlook = sunny |Class = yes) * P(Temp = cool |Class = yes) * P(Humidity = high |Class = yes) * P(Windy = true |Class = yes) * P(Class = yes)/P(evidence)
= *2/9 * 3/9 * 3/9 * 3/9 * 9/14 /P(evidence)*
= *0.0053 /P(evidence)*

- Compute probability of Class=no:

P(Class = no | Outlook = sunny & Temp = cool & Humidity = high & Windy = true)
= ???

Exercise!

How to Make Prediction?

What class does this new example belong to? Class=yes OR Class=no?

Outlook	Temp	Humidity	Windy	Class
sunny	cool	high	true	?

P(yes|attributes) = 0.0053/P(evidence)

P(no|attributes) = 0.0206/P(evidence)

$$\text{P(evidence)} = P(X_1 = a_1|Y = 1) \times \cdots \times P(A_k = a_k|Y = 1) \times P(Y = 1)$$
$$+P(X_1 = a_1|Y = 0) \times \cdots \times P(A_k = a_k|Y = 0) \times P(Y = 0)$$

After normalizing it, the class probabilities become:

P(yes|attributes) = 0.0053 / (0.0053 + 0.0206) = 0.205

P(no|attributes) = 0.0206 / (0.0053 + 0.0206) = 0.795

Naïve Bayes Summary: Strengths

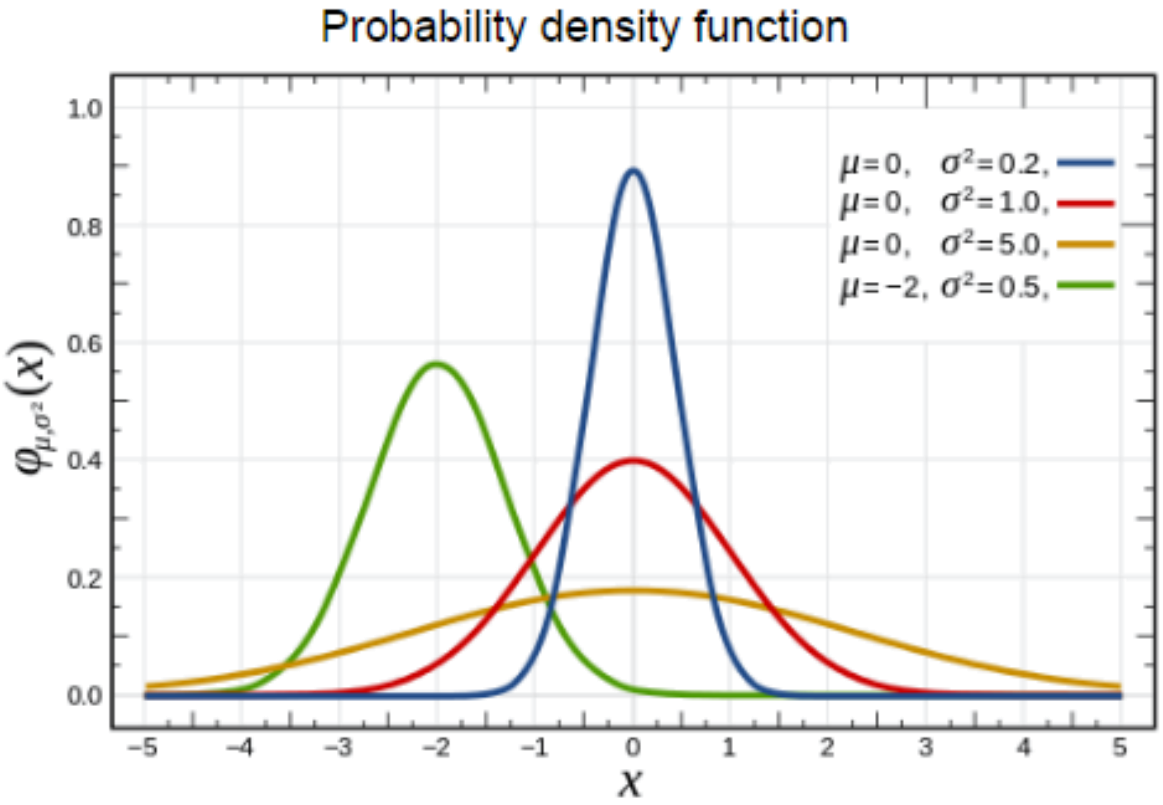
- Strengths
  - Simple; efficient in both storage space and computation time
  - Performs well for classification in many real-world applications
  - Incremental learning (no need to reprocess all past training data when new data become available)
  - Easily handles missing values (in training, ignore missing values when doing frequency count; when testing, the attribute with missing value is not included in probability estimate calculation)
  - Naturally handle multi-class classification problem

Naïve Bayes Summary: Weaknesses

- Weaknesses
  - Cannot naturally handle numerical feature (needs special care)
  - Independence assumption. In reality, many attributes have obvious correlations, such as age and income in demographic data.
  - Zero-frequency problem: if you have no occurrences of a class label and a certain attribute value, the frequency-based probability and thus the posterior probability estimate will be zero. (Solution: Laplace smoothing)

(Optional) Handling Continuous Attributes

- Impose more assumptions like ***Gaussian distribution***, i. e. , $f_k(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2}(\frac{x-\mu_k}{\sigma})^2)$



(Optional) Zero-Frequency Problem

- Zero-frequency problem: in a rare event the frequency is 0, and thus the whole P(Y=1|..) will be zero. Thus we want to assign small probability to rare event. The method is called Laplace smoothing.

$$P(X_1 = a_1 | Y = 1) = \frac{count(X_1 = a_1 \; and \; Y = 1) + \alpha lpha}{count(Y = 1) + |X_1| * \alpha lpha}$$

Number of unique values of attribute X_1

** alpha is the Laplace smoothing parameter.

Lab: Naïve Bayes Classifier

- In Python sklearn, there are three versions of Naïve Bayes classifiers.
  - BernoulliNB (more suitable for binary predictor features)
  - MultinomialNB (more suitable for discrete predictor features)
  - GaussianNB (more suitable for continuous predictor features)
- Files needed
  - Naïve Bayes.ipynb (Python file)
  - UniversalBank.csv (dataset)