Probability Background

Probability Theory

- What is probability?
  - Probability is the measure of the likelihood [0,1] that an event will occur
  - Given a fair coin (both sides are equally likely to turn up)
    - P(heads) = 1 – P(tails) = 0.5
  - Given a fair die
    - P(getting a 6) = 1/6
    - P(getting a 5) = 1/6
    - P(getting an 11) from two dice?

Conditional Probability

- What is the probability that a student is female?
  - P(S='female') Here s is the random variable and 'female' is the value
  - P(S='female') = 0.5 without any additional knowledge
  - But can we do better if know that the student is studying Computer Science at UiS?
    - P(S='female' | S is an CS student at UiS)
    - If we know there are 1% female students at UiS studying CS
    - P(S='female' | S is an CS student at UiS) = 0.01

Conditional Probability

Name	Gender	Study
Tom	M	CS
Harry	M	CS
Annika	F	Medicine
Kate	F	CS
Ingrid	F	Medicine
Anne	F	Medicine
Christian	M	CS
Bent	M	Sociology
Helle	F	Sociology
Jens	M	Sociology

- P(F) $= \frac{5}{10}$ = 0.5
- P(F | CS) $= \frac{1}{4} = 0.25$
- P(F | CS) $= \frac{P(F \cap CS)}{P(CS)} = \frac{1/10}{4/10} =$ 0.25

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

Bayes Theorem

- $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$
- $P(B \mid A) = \frac{P(A \cap B)}{P(A)} \Rightarrow P(A \cap B) = P(B \mid A)P(A)$
- $P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$
- For example, what is the probability that a female student will choose to study computer science?
  - $P(CS \mid F) = \frac{P(F \mid CS)P(CS)}{P(F)} = \frac{0.25 * 0.4}{0.5}$

Name	Gender	Study
Tom	M	CS
Harry	M	CS
Annika	F	Medicine
Kate	F	CS
Ingrid	F	Medicine
Anne	F	Medicine
Christian	M	CS
Bent	M	Sociology
Helle	F	Sociology
Jens	M	Sociology

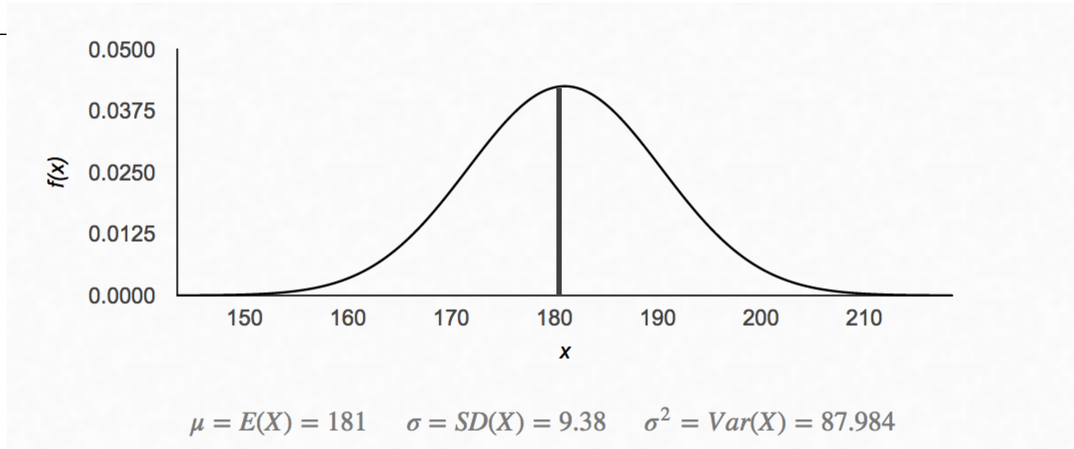
For Continuous Values

- How to find conditional probability of a continuous variable?
- How to find P(H=180)?
- Assuming the height distribution follows normal distribution (bell curve)
- We can use Gaussian distribution to compute this

Name	Gender	Study	Height (cm)
Tom	M	CS	180
Harry	M	CS	175
Annika	F	Medicine	165
Kate	F	CS	200
Ingrid	F	Medicine	184
Anne	F	Medicine	178
Christian	M	CS	185
Bent	M	Sociology	179
Helle	F	Sociology	175
Jens	M	Sociology	189

Gaussian/Normal Distribution

Name	Gender	Height (cm)
Tom	M	180
Harry	M	175
Annika	F	165
Kate	F	200
Ingrid	F	184
Anne	F	178
Christian	M	185
Bent	M	179
Helle	F	175
Jens	M	189



- Mean ($\boldsymbol{\mu}$) = $\frac{180+175+165+200+184+178+185+179+175+189}{10} = 181$
- Standard deviation (σ) = $\sqrt[2]{\frac{\sum(x_i - \boldsymbol{\mu})^2}{N}} = 9.38$
- P(H=180) = $\frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi * 9.38}} e^{-\frac{(180-181)^2}{2*9.38^2}} = 0.216 * 0.994 = 0.214$

Conditional Probability for Continuous Variables

- How to find P(H=180 | F)? What is the probability of a female student being 180 cm tall?
  - We follow same Gaussian distribution

$$\frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

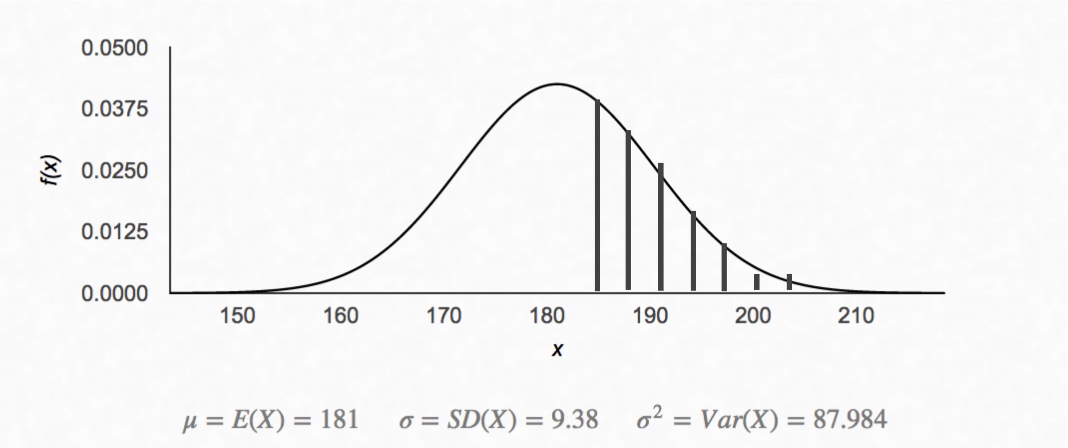
  - Except we compute μ_F and σ_F which are mean and standard deviation for Female students only

Name	Gender	Height (cm)
Tom	M	180
Harry	M	175
Annika	F	165
Kate	F	200
Ingrid	F	184
Anne	F	178
Christian	M	185
Bent	M	179
Helle	F	175
Jens	M	189

Conditional Probability for Continuous Range Values

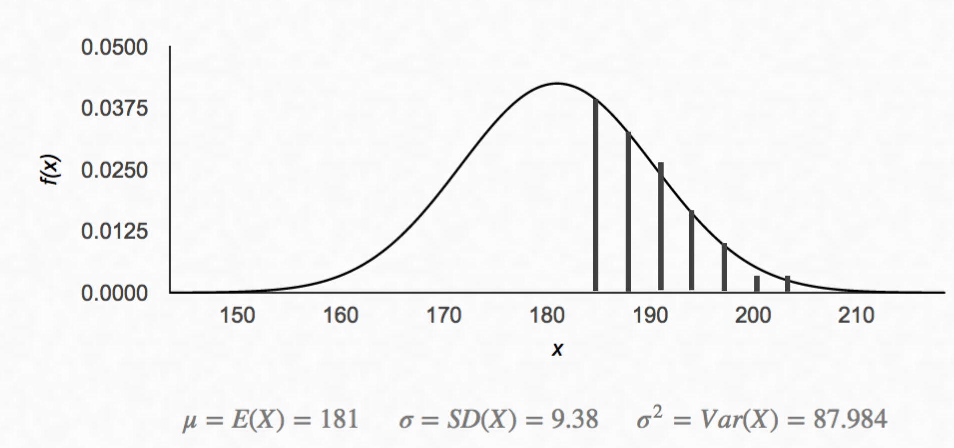
- How to find P(H >185)? What is the probability of a student being more than 180 cm tall?

- $$\frac{1}{\sqrt{2\pi\sigma}}\int_{181}^{+\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Probability Density Function

- P(H) is also called a probability density function in general
- Any h ∈ H is a value between 0 and 1
- $P_{data}(X)$ is true distribution
- $P_{model}(X)$ is an estimate over true distribution



Parametric Modeling

- There may be many $P_{model}(X)$ that describe the data best.
- We want to find best set of parameters $\boldsymbol{\theta}$ so that best describes $P_{data}(X)$
- Likelihood
  - $L(\boldsymbol{\theta} \mid x)$ x is fixed, for each $\boldsymbol{\theta}$ how consistent the model is.
  - $L(\boldsymbol{\theta} \mid x) = \prod_{x \in X} P_{\boldsymbol{\theta}}(X) = \sum_{x \in X} log P_{\boldsymbol{\theta}}(X)$

Maximum Likelihood Estimation

- Goal is to find optimal $\boldsymbol{\theta}$ that maximizes $L(\boldsymbol{\theta} \mid x)$
- $\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \operatorname{argmax} L(\boldsymbol{\theta} \mid x)$
- Most ML models minimize loss function so they minimize negative log likelihood.

$$\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \operatorname{argmin} -log P_{\boldsymbol{\theta}}(X)$$

Bayes Classifier

- A probabilistic framework for solving classification problems
- Conditional Probability:

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

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

- Bayes theorem:

$$P(C \mid A) = \frac{P(A \mid C)P(C)}{P(A)}$$

Example of Bayes Theorem

- Given:
  - A doctor knows that meningitis causes stiff neck 50% of the time
  - Prior probability of any patient having meningitis is 1/50,000
  - Prior probability of any patient having stiff neck is 1/20

- If a patient has stiff neck, what's the probability he/she has meningitis?

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

Another Example

- Assume you are diagnosed with a terrible disease which only affects **0.1%** of the population.
- Your doctor says the test misdiagnoses in **1%** of the cases.
- What is the probability (0 to 1) that you actually have this disease?

Bayesian Classifiers

- Consider each attribute and class label as random variables

- Given a record with attributes ($A_1, A_2, \ldots, A_n$)
  - Goal is to predict class C
  - Specifically, we want to find the value of C that maximizes $P(C|A_1, A_2, \ldots, A_n)$

- Can we estimate $P(C|A_1, A_2, \ldots, A_n)$ directly from data?

Bayesian Classifiers

- Approach:
  - compute the posterior probability $P(C \mid A_1, A_2, \ldots, A_n)$ for all values of C using the Bayes theorem

$$\boxed{P(C \mid A_1 A_2 \ldots A_n) = \frac{P(A_1 A_2 \ldots A_n \mid C) P(C)}{P(A_1 A_2 \ldots A_n)}}$$

  - Choose value of C that maximizes
    $P(C \mid A_1, A_2, \ldots, A_n)$
  - Equivalent to choosing value of C that maximizes
    $P(A_1, A_2, \ldots, A_n \mid C)\, P(C)$
- How to estimate $P(A_1, A_2, \ldots, A_n \mid C)$?

Naïve Bayes Classifier

- Assume independence among attributes A_i when class is given:
  - $P(A_1, A_2, \ldots, A_n | C) = P(A_1 | C_j)\, P(A_2 | C_j) \ldots P(A_n | C_j)$
  - Can estimate $P(A_i | C_j)$ for all A_i and C_j.
  - New point is classified to C_j if $P(C_j) \prod P(A_i | C_j)$ is maximal.

How to Estimate Probabilities from Data?

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- Class: $P(C) = N_C/N$
  - e.g., $P(No) = 7/10$,
    $P(Yes) = 3/10$

- For discrete attributes:

$$P(A_i \mid C_k) = |A_{ik}| / N_c$$

  - where $|A_{ik}|$ is number of instances having attribute A_i and belongs to class C_k
  - Examples:

P(Status=Married|No) = 4/7
P(Refund=Yes|Yes)=0

How to Estimate Probabilities from Data?

- For continuous attributes:
  - Discretize the range into bins
    - one ordinal attribute per bin
    - violates independence assumption
  - Two-way split: (A < v) or (A > v)
    - choose only one of the two splits as new attribute
  - Probability density estimation:
    - Assume attribute follows a normal distribution
    - Use data to estimate parameters of distribution (e.g., mean and standard deviation)
    - Once probability distribution is known, can use it to estimate the conditional probability $P(A_i|c)$

k

How to Estimate Probabilities from Data?

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- Normal distribution:

$$P(A_i \mid c_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(A_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

  - One for each (A_i, c_i) pair

- For (Income, Class=No):
  - If Class=No
    - sample mean = 110
    - sample variance = 2975

$$P(Income = 120 \mid No) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

Example of Naïve Bayes Classifier

Given a Test Record:

$$X = (\text{Refund} = \text{No}, \text{Married}, \text{Income} = 120\text{K})$$

naive Bayes Classifier:

P(Refund=Yes|No) = 3/7
P(Refund=No|No) = 4/7
P(Refund=Yes|Yes) = 0
P(Refund=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No)=1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/7
P(Marital Status=Divorced|Yes)=1/7
P(Marital Status=Married|Yes) = 0

For taxable income:
If class=No: sample mean=110
sample variance=2975
If class=Yes: sample mean=90
sample variance=25

- P(X|Class=No) = P(Refund=No|Class=No) $\times$ P(Married| Class=No) $\times$ P(Income=120K| Class=No) = 4/7 $\times$ 4/7 $\times$ 0.0072 = 0.0024

- P(X|Class=Yes) = P(Refund=No| Class=Yes) $\times$ P(Married| Class=Yes) $\times$ P(Income=120K| Class=Yes) = 1 $\times$ 0 $\times$ 1.2 $\times$ 10^{-9} = 0

Since P(X|No)P(No) > P(X|Yes)P(Yes)

Therefore P(No|X) > P(Yes|X)

=> Class = No

Naïve Bayes Classifier

- If one of the conditional probability is zero, then the entire expression becomes zero
- Probability estimation:

$$\text{Original}: P(A_i \mid C) = \frac{N_{ic}}{N_c}$$

$$\text{Laplace}: P(A_i \mid C) = \frac{N_{ic} + 1}{N_c + c}$$

$$\text{m - estimate}: P(A_i \mid C) = \frac{N_{ic} + mp}{N_c + m}$$

c: number of classes

p: prior probability

m: parameter

Example of Naïve Bayes Classifier

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

A: attributes

M: mammals

N: non-mammals

$$P(A|M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A|N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A|M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A|N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

P(A|M)P(M) > P(A|N)P(N)

=> Mammals

Probability Density Function

- P(D I E) is simplified to P(D).
- P(HD I E, D) cannot be simplified.
- P(Hb I HD,E,D) is simplified to P(HbID)
- P(CP I Hb, H D, E, D) is simplified to P(C PIHb, H D)
- P(BP I CP,Hb,HD,E,D) is simplified to P(BP I HD).

Naïve Bayes (Summary)

- Robust to isolated noise points
- Handle missing values by ignoring the instance during probability estimate calculations
- Robust to irrelevant attributes
- Independence assumption may not hold for some attributes
  - Use other techniques such as Bayesian Belief Networks (BBN)

Literature

- Chapter 6 from the Tan et. al. Textbook.