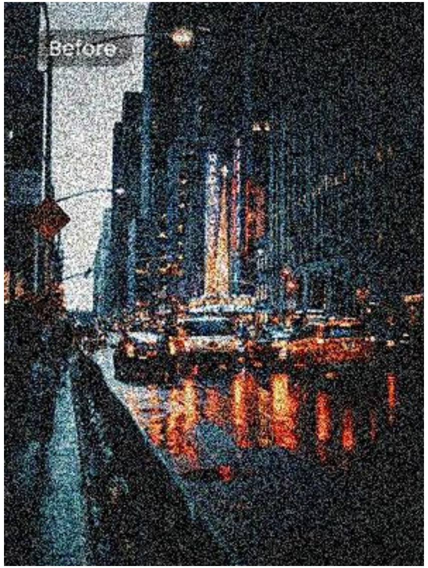
# Intro to Probability

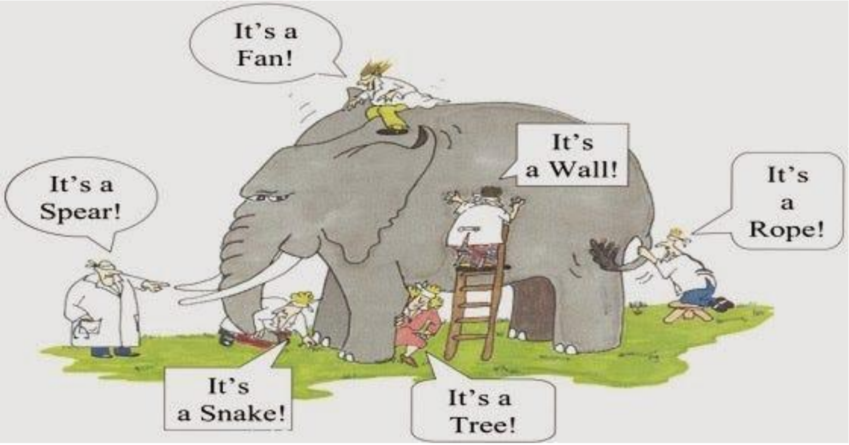
Mahesh Mohan M R
Centre of Excellence in AI
Indian Institute of Technology Kharagpur

# Probability

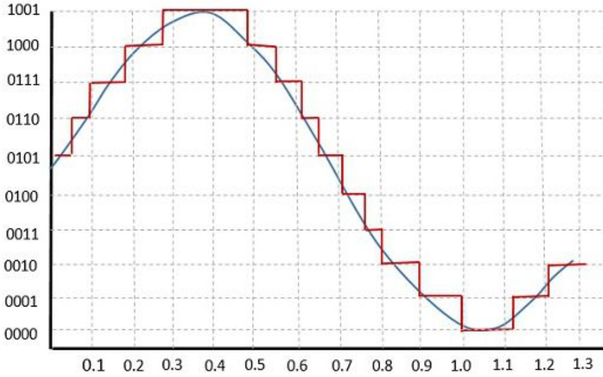
- Solving machine learning problems requires to deal with uncertain quantities, as well as with stochastic (non-deterministic) quantities
  - Probability theory provides a mathematical framework for representing and quantifying uncertain quantities
- There are different sources of uncertainty:



Inherent stochasticity



Incomplete observability



Incomplete modeling

# Probability

- Intuition:
  - In a process, several outcomes are possible
  - When the process is repeated a large number of times, each outcome occurs with a ***relative frequency***, or ***probability***
  - If a particular outcome occurs more often, we say it is more probable
- Probability arises in two contexts
  - In actual repeated experiments
    - Example: You record the color of 1,000 cars driving by. 57 of them are green. You estimate the probability of a car being green as 57/1,000 = 0.057.
  - In idealized conceptions of a repeated process
    - Example: You consider the behavior of an unbiased six-sided die. The expected probability of rolling a 5 is 1/6 = 0.1667.
    - Example: You need a model for how people's heights are distributed.

# Random Variable

- A ***random variable*** $X$ is a variable that can take on different values
  - Example: $X$ = rolling a die
    - Possible values of $X$ comprise the **sample space**, or **outcome space**, $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$
    - We denote the event of "seeing a 5" as $\{X = 5\}$ or $X = 5$
    - The probability of the event is $P(\{X = 5\})$ or $P(X = 5)$
    - Also, $P(5)$ can be used to denote the probability that $X$ takes the value of 5
- A ***probability distribution*** is a description of how likely a random variable is to take on each of its possible states
  - A compact notation is common, where $P(X)$ is the probability distribution over the random variable $X$
    - Also, the notation $\mathrm{X} \sim P(X)$ can be used to denote that the random variable $X$ has probability distribution $P(X)$
- Random variables can be discrete or continuous
  - Discrete random variables have finite number of states: e.g., the sides of a die
  - Continuous random variables have infinite number of states: e.g., the height of a person

# Axioms of probability

- The probability of an event $\mathcal{A}$ in the given sample space $\mathcal{S}$, denoted as $P(\mathcal{A})$, must satisfies the following properties:
  - Non-negativity
    - For any event $\mathcal{A} \in \mathcal{S}$, $P(\mathcal{A}) \geq 0$
  - All possible outcomes
    - Probability of the entire sample space is 1, $P(\mathcal{S}) = 1$
  - Additivity of disjoint events
    - For all events $\mathcal{A}_1, \mathcal{A}_2 \in \mathcal{S}$ that are mutually exclusive $(\mathcal{A}_1 \cap \mathcal{A}_2 = \emptyset)$, the probability that both events happen is equal to the sum of their individual probabilities, $P(\mathcal{A}_1 \cup \mathcal{A}_2) = P(\mathcal{A}_1)+P(\mathcal{A}_2)$

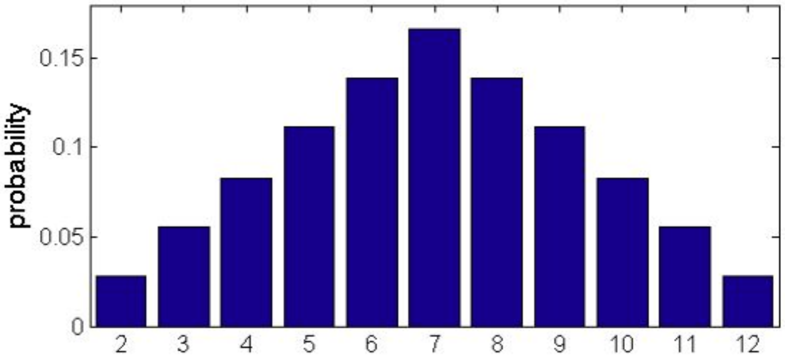
- The probability of a random variable $P(X)$ must obey the axioms of probability over the possible values in the sample space $\mathcal{S}$

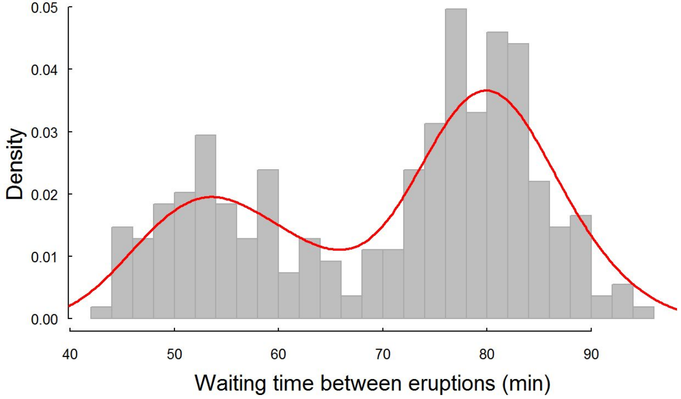
# Probability Functions

- A probability distribution over discrete variables may be described using a ***probability mass function*** (PMF)
  - E.g., sum of two dice

- A probability distribution over continuous variables may be described using a ***probability density function*** (PDF)
  - E.g., waiting time between eruptions of Old Faithful
  - A PDF gives the probability of an infinitesimal region with volume $\delta X$
  - To find the probability over an interval $[a, b]$, we can integrate the PDF as follows:

$$P(X \in [a,b]) = \int_a^b P(X)dX$$





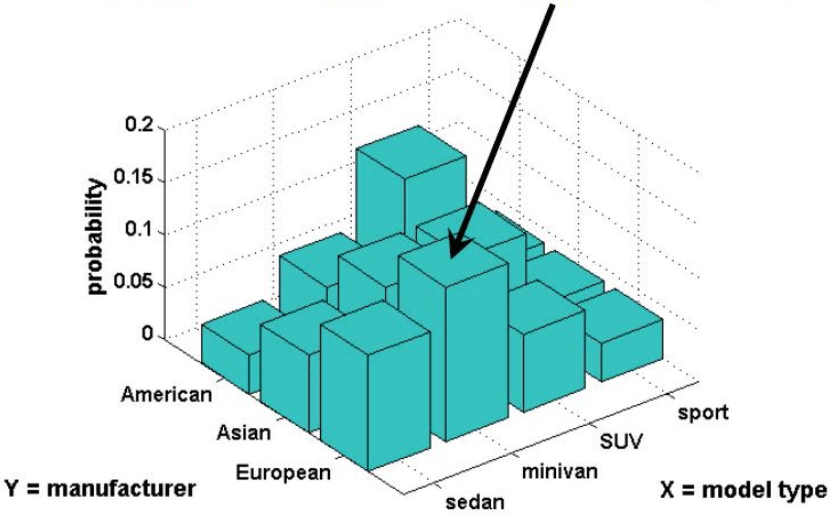
# Multivariate Random Variables

- We may need to consider several random variables at a time
  - If several random processes occur in parallel or in sequence
  - E.g., to model the relationship between several diseases and symptoms
  - E.g., to process images with millions of pixels (each pixel is one random variable)
- Next, we will study probability distributions defined over multiple random variables
  - These include joint, conditional, and marginal probability distributions
- The individual random variables can also be grouped together into a random vector, because they represent different properties of an individual statistical unit
- A *multivariate random variable* is a vector of multiple random variables $\mathbf{X} = (X_1, X_2, \dots, X_n)^T$

# Joint Probability Distribution

- Probability distribution that acts on many variables at the same time is known as a ***joint probability distribution***
- Given any values $x$ and $y$ of two random variables $X$ and $Y$, what is the probability that $X = x$ and $Y = y$ simultaneously?
  - $P(X = x, Y = y)$ denotes the joint probability
  - We may also write $P(x, y)$ for brevity

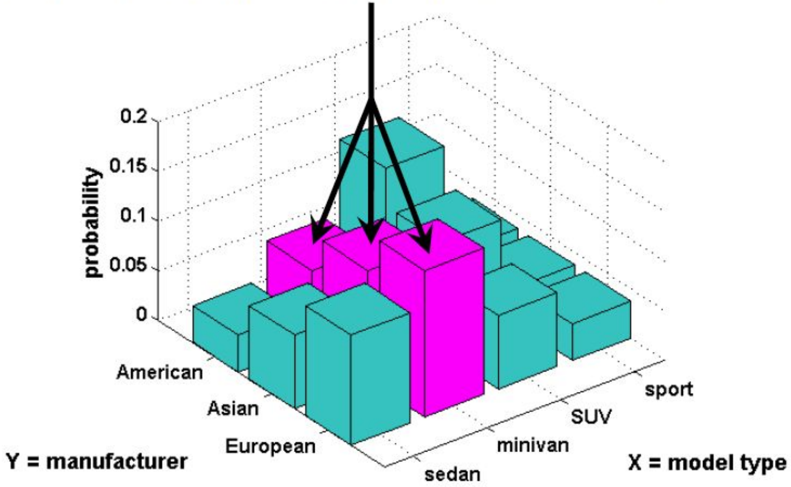
joint probability: $p($ $X$ = minivan, $Y$ = European $)$ = 0.1481



# Marginal Probability Distribution

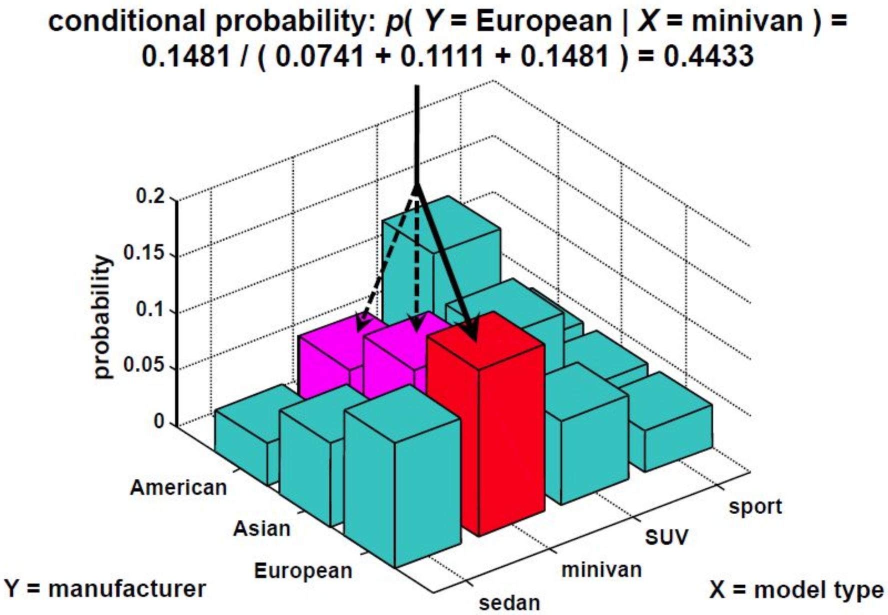
- ***Marginal probability distribution*** is the probability distribution of a single variable
  - It is calculated based on the joint probability distribution $P(X, Y)$
  - I.e., using the sum rule: $P(X = x) = \sum_y P(X = x, Y = y)$
    - For continuous random variables, the summation is replaced with integration, $P(X = x) = \int P(X = x, Y = y)\, dy$
  - This process is called marginalization

marginal probability: $p($ $X$ = minivan $)$ = 0.0741 + 0.1111 + 0.1481 = 0.3333



# Conditional Probability Distribution

- ***Conditional probability distribution*** is the probability distribution of one variable provided that another variable has taken a certain value
  - Denoted $P(X = x|\, Y = y)$
- Note that: $P(X = x|\, Y = y) = \frac{P(X=x, Y=y)}{P(Y=y)}$



# Bayes' Theorem

- *Bayes' theorem* – allows to calculate conditional probabilities for one variable when conditional probabilities for another variable are known

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

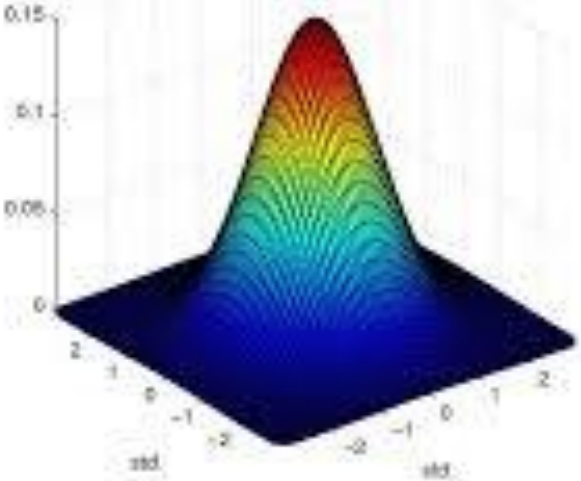
- Also known as Bayes' rule
- Multiplication rule for the joint distribution is used: $P(X,Y) = P(Y|X)P(X)$
- By symmetry, we also have: $P(Y,X) = P(X|Y)P(Y)$

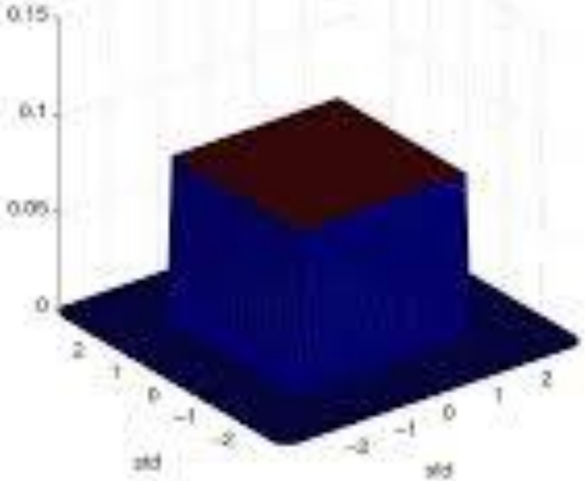
# Independence

- Two random variables $X$ and $Y$ are ***independent*** if the occurrence of $Y$ does not reveal any information about the occurrence of $X$
  - E.g., two successive rolls of a die are independent
- Therefore, we can write: $P(X|Y) = P(X)$
  - The following notation is used: $X \perp Y$
  - Also note that for independent random variables: $P(X, Y) = P(X)P(Y)$
- In all other cases, the random variables are ***dependent***
  - Getting a king on successive draws form a deck (the drawn card is not replaced)
- Two random variables $X$ and $Y$ are ***conditionally independent*** given another random variable $Z$ if and only if $P(X, Y|Z) = P(X|Z)P(Y|Z)$
  - This is denoted as $X \perp Y|Z$

# Continuous Multivariate Distributions

- Same concepts of joint, marginal, and conditional probabilities apply for continuous random variables
- The probability distributions use integration of continuous random variables, instead of summation of discrete random variables
  - Example:  Gaussian/Uniform probability distribution in two dimensions





# Expected Value

- The *expected value* or *expectation* of a function $f(X)$ with respect to a probabilit distribution $P(X)$ is the average (mean) when $X$ is drawn from $P(X)$
- For a discrete random variable $X$, it is calculated as

$$\mathbb{E}_{X\sim P}[f(X)] = \sum_{X} P(X)f(X)$$

- For a continuous random variable $X$, it is calculated as

$$\mathbb{E}_{X\sim P}[f(X)] = \int P(X)f(X)\,dX$$

  - When the identity of the distribution is clear from the context, we can write $\mathbb{E}_X[f(X)]$
  - If it is clear which random variable is used, we can write just $\mathbb{E}[f(X)]$
- Mean is the most common measure of central tendency of a distribution
  - For a random variable: $f(X_i) = X_i \quad \Rightarrow \quad \mu = \mathbb{E}[X_i] = \sum_i P(X_i) \cdot X_i$
  - This is similar to the mean of a sample of observations: $\mu = \frac{1}{N}\sum_i X_i$
  - Other measures of central tendency: median, mode

# Variance

- ***Variance*** gives the measure of how much the values of the function $f(X)$ de from the expected value as we sample values of X from $P(X)$

$$\mathrm{Var}\big(f(X)\big) = \mathbb{E}[(f(X) - \mathbb{E}[f(X)])^2]$$

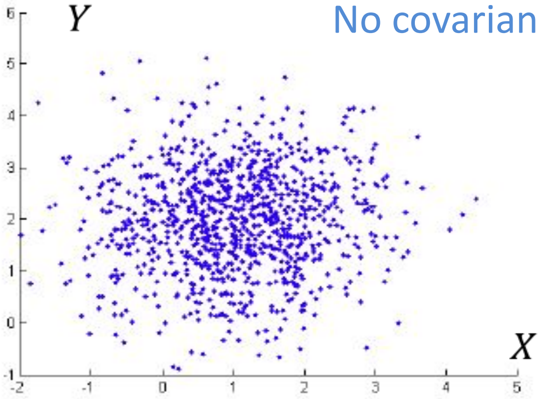
- When the variance is low, the values of $f(X)$ cluster near the expected value
- Variance is commonly denoted with $\sigma^2$
  - The above equation is similar to a function $f(X_i) = X_i - \mu$
  - We have $\sigma^2 = \sum_i P(X_i) \cdot (X_i - \mu)^2$
  - This is similar to the formula for calculating the variance of a sample of observati $\sigma^2 = \frac{1}{N-1} \sum_i (X_i - \mu)^2$
- The square root of the variance is the ***standard deviation***
  - Denoted $\sigma = \sqrt{\mathrm{Var}(X)}$

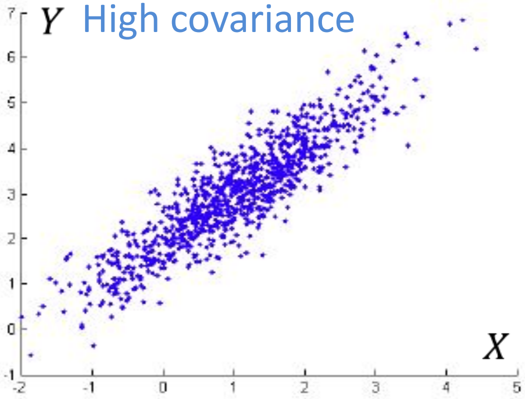
# Covariance

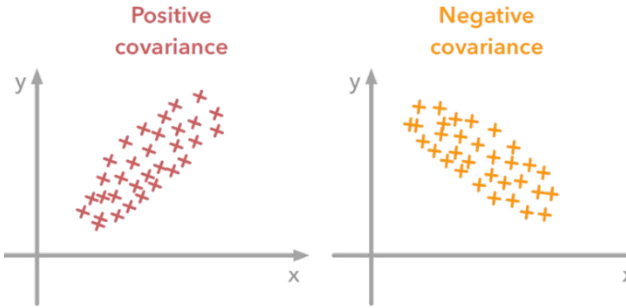
- *Covariance* gives the measure of how much two random variables are linearly related to each other

$$\mathrm{Cov}\big(f(X), g(Y)\big) = \mathbb{E}[(f(X) - \mathbb{E}[f(X)])(g(Y) - \mathbb{E}[g(Y)])]$$

- If $f(X_i) = X_i - \mu_X$ and $g(Y_i) = Y_i - \mu_Y$
  - Then, the covariance is: $\mathrm{Cov}(X,Y) = \sum_i P(X_i, Y_i) \cdot (X_i - \mu_X) \cdot (Y_i - \mu_Y)$
  - Compare to covariance of actual samples: $\mathrm{Cov}(X,Y) = \frac{1}{N-1} \sum_i (Y_i - \mu_X)(Y_i - \mu_Y)$
- The covariance measures the tendency for $X$ and $Y$ to deviate from their means in same (or opposite) directions at same time







# Covariance Matrix

- *Covariance matrix* of a multivariate random variable $\mathbf{X}$ with states $\mathbf{x} \in \mathbb{R}^{\boldsymbol{n}}$ is an $n \times n$ matrix, such that

$$\mathrm{Cov}(\mathbf{X})_{i,j} = \mathrm{Cov}\left(\mathbf{x}_i, \mathbf{x}_j\right)$$

- I.e.,

$$\mathrm{Cov}(\mathbf{X}) = \begin{bmatrix} \mathrm{Cov}(\mathbf{x}_1, \mathbf{x}_1) & \mathrm{Cov}(\mathbf{x}_1, \mathbf{x}_2) & \cdots & \mathrm{Cov}(\mathbf{x}_1, \mathbf{x}_{\boldsymbol{n}}) \\ \mathrm{Cov}(\mathbf{x}_2, \mathbf{x}_1) & \ddots & & \mathrm{Cov}(\mathbf{x}_2, \mathbf{x}_{\boldsymbol{n}}) \\ \vdots & & & \vdots \\ \mathrm{Cov}(\mathbf{x}_n, \mathbf{x}_1) & \mathrm{Cov}(\mathbf{x}_n, \mathbf{x}_2) & \cdots & \mathrm{Cov}(\mathbf{x}_{\boldsymbol{n}}, \mathbf{x}_{\boldsymbol{n}}) \end{bmatrix}$$

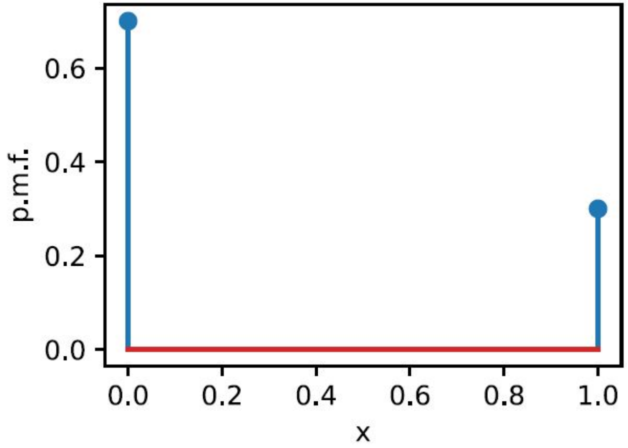
- The diagonal elements of the covariance matrix are the variances of the elements of the vector

$$\mathrm{Cov}(\mathbf{x}_i, \mathbf{x}_i) = \mathrm{Var}(\mathbf{x}_i)$$

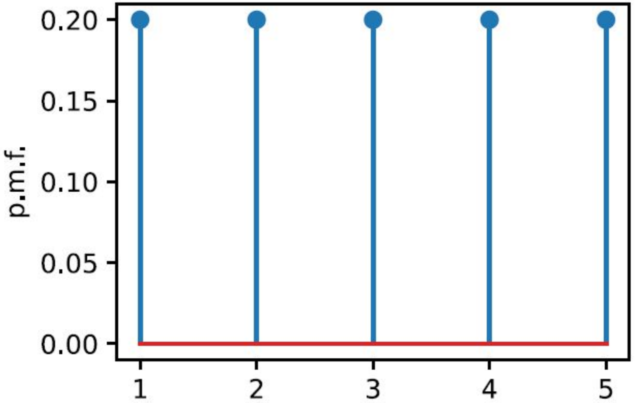
- Also note that the covariance matrix is symmetric, since $\mathrm{Cov}\left(\mathbf{x}_i, \mathbf{x}_j\right) = \mathrm{Cov}\left(\mathbf{x}_j, \mathbf{x}_i\right)$

# Probability Distributions

- ***Bernoulli distribution***
  - Binary random variable $X$ with states $\{0, 1\}$
  - The random variable can encodes a coin flip which comes up 1 with probability $p$ and 0 with probability $1 - p$
  - Notation: $X \sim Bernoulli(p)$



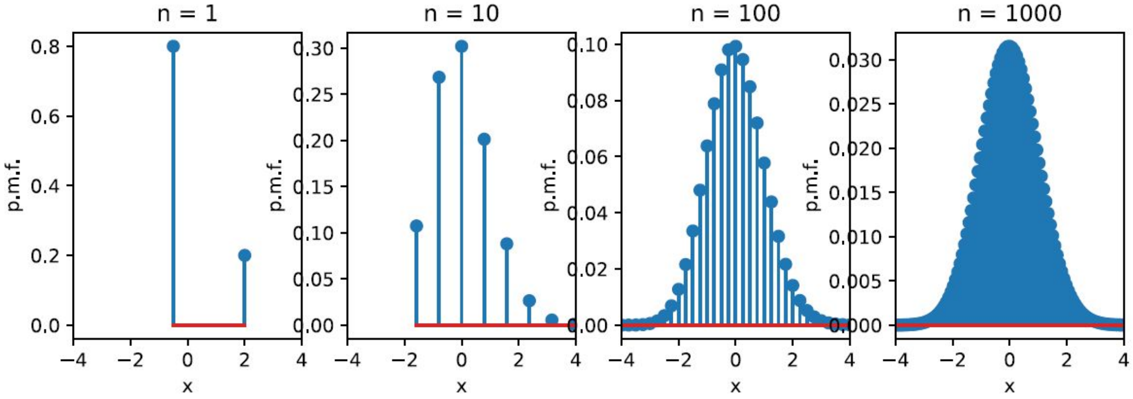
- ***Uniform distribution***
  - The probability of each value $i \in \{1,2, \dots, n\}$ is $p_i = \frac{1}{n}$
  - Notation: $X \sim U(n)$
  - Figure: $n = 5,\ p = 0.2$

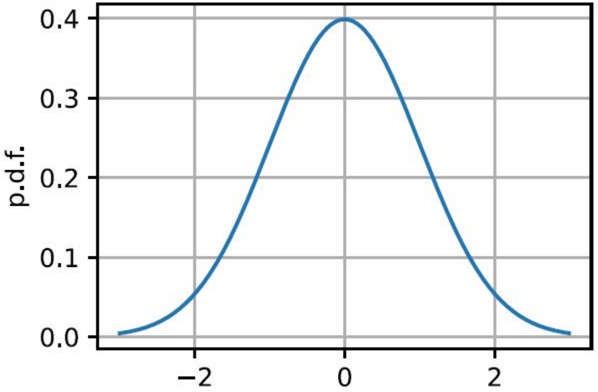


# Probability Distributions

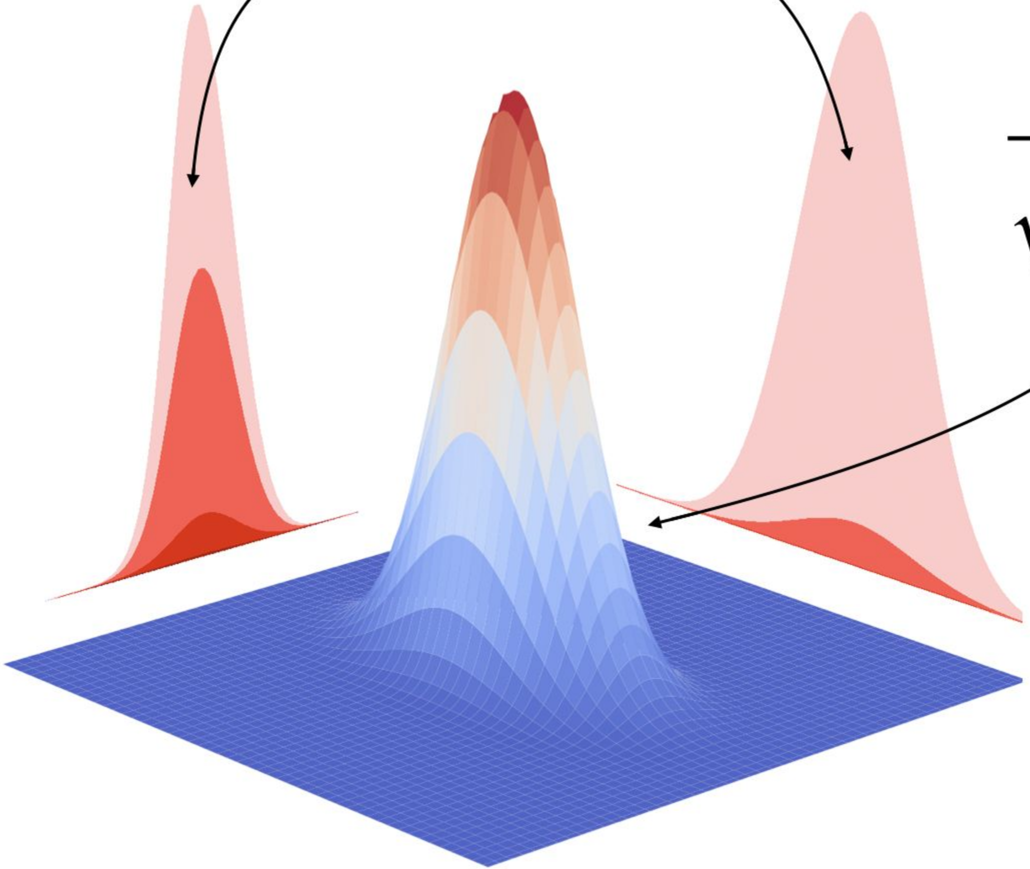
- ***Gaussian distribution***
  - The most well-studied distribution
    - Referred to as normal distribution or informally bell-shaped distribution
  - Defined with the mean $\mu$ and variance $\sigma^2$
  - Notation: $X \sim \mathcal{N}(\mu, \sigma^2)$
  - For a random variable $X$ with $n$ independent measurements, the density is

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





Univariate Gaussians



Multivariate Gaussian

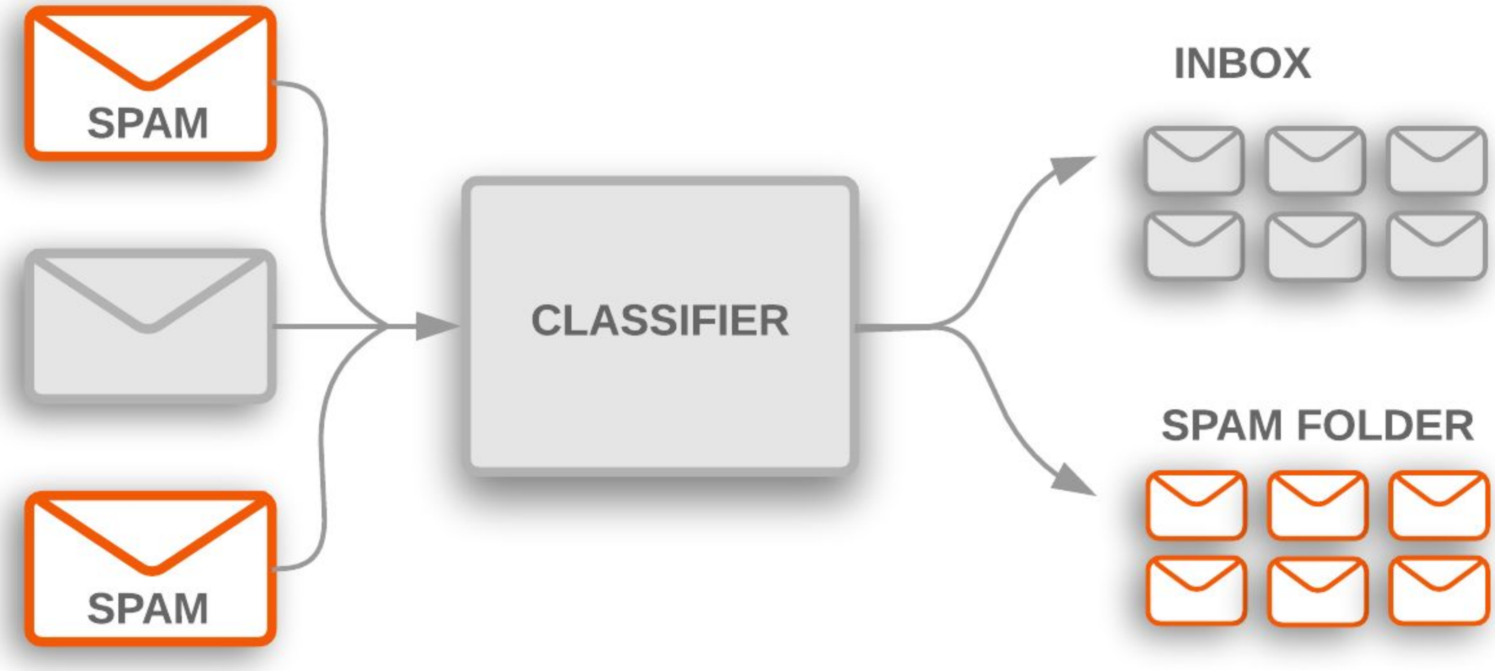
$$\frac{1}{\sqrt{(2\pi)^2\,|\,\hat{\Sigma}\,|}}\,e^{-\frac{1}{2}(\vec{r}-\overrightarrow{\mu})^T\hat{\Sigma}^{-1}(\vec{r}-\overrightarrow{\mu})}$$

$$\hat{\Sigma} = \begin{pmatrix} \sigma_x^2 & cov(x,y) \\ cov(y,x) & \sigma_y^2 \end{pmatrix}$$

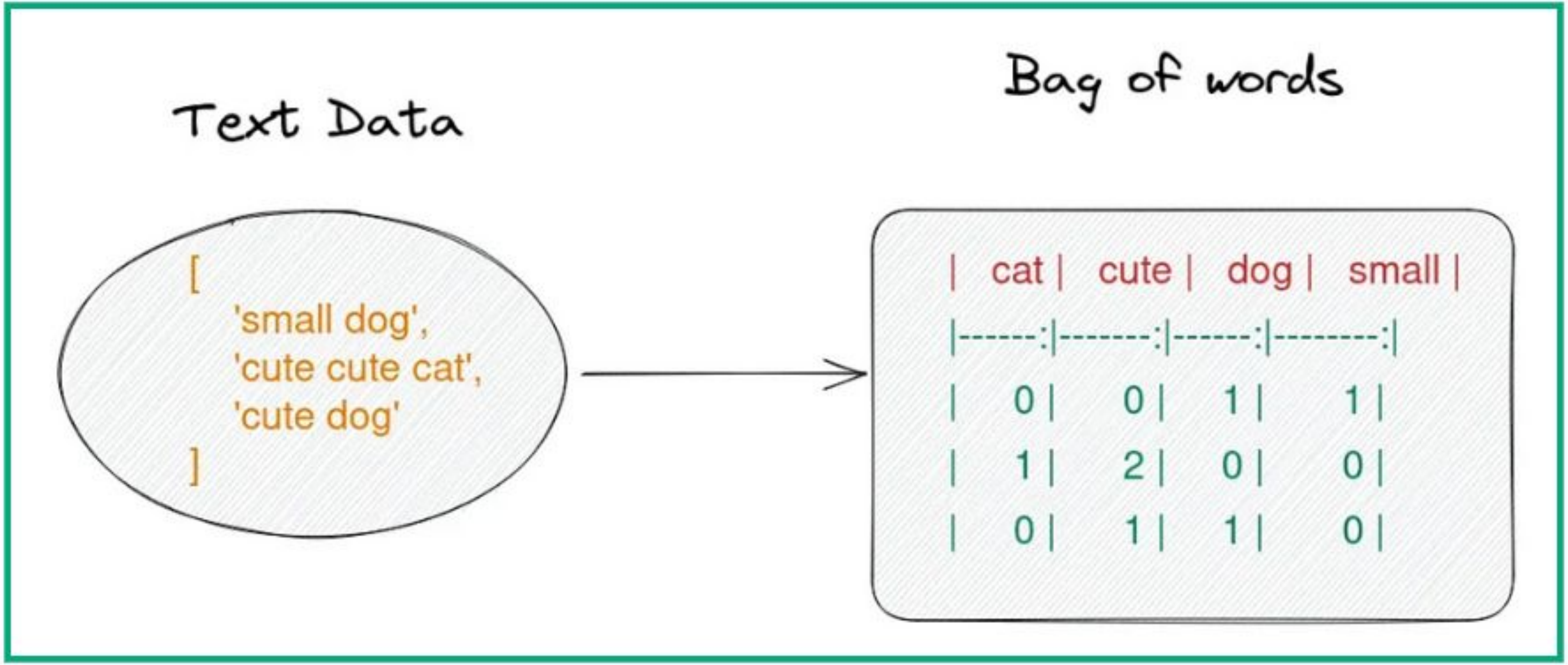
# Things We'd Like to Do

- Spam Classification
  - Given an email, predict whether it is spam or not
- Medical Diagnosis
  - Given a list of symptoms, predict whether a patient has disease X or not
- Weather
  - Based on temperature, humidity, etc… predict if it will rain tomorrow

# Practical Example: Spam or not



# Practical Example: Spam or not



```
[
    'small dog',
    'cute cute cat',
    'cute dog'
]
```

| cat | cute | dog | small |
|-----:|-------:|------:|--------:|
| 0 | 0 | 1 | 1 |
| 1 | 2 | 0 | 0 |
| 0 | 1 | 1 | 0 |

# Practical Example: Spam or not

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

fairy
always
love
to
it
it
whimsical
it
I
and
seen
are
anyone
friend
happy
dialogue
adventure
recommend
who
sweet
of
satirical
movie
it
it
I
but
to
romantic
I
several
yet
again
it
the
humor
the
seen
would
to
scenes
I
the
manages
fun
I
and
the
times
and
about
while
whenever
have
conventions
with

| | |
|---|---|
| it | 6 |
| I | 5 |
| the | 4 |
| to | 3 |
| and | 3 |
| seen | 2 |
| yet | 1 |
| would | 1 |
| whimsical | 1 |
| times | 1 |
| sweet | 1 |
| satirical | 1 |
| adventure | 1 |
| genre | 1 |
| fairy | 1 |
| humor | 1 |
| have | 1 |
| great | 1 |
| ... | ... |

# The Bayes Classifier

- A good strategy is to predict:

$$\arg\max_{Y} P(Y|X_1, \ldots, X_n)$$

  - (for example: what is the probability that it is spam given n dimension Bag of words features?)

- So … How do we compute that?

- Use Bayes Rule!

$$P(Y|X_1,\ldots,X_n) = \frac{\overset{\text{Likelihood}}{\downarrow}{P(X_1,\ldots,X_n|Y)}\,\overset{\text{Prior}}{\downarrow}{P(Y)}}{\underset{\substack{\uparrow\\ \text{Normalization Constant}}}{P(X_1,\ldots,X_n)}}$$

- Why did this help? Well, we think that we might be able to specify how features are “generated” by the class label

- Let’s expand this for our SPAm Classification task (Spam:5 and Non-spam:6):

$$
\begin{aligned}
P(Y=5|X_1,\ldots,X_n) &= \frac{P(X_1,\ldots,X_n|Y=5)P(Y=5)}{P(X_1,\ldots,X_n|Y=5)P(Y=5)+P(X_1,\ldots,X_n|Y=6)P(Y=6)} \\
P(Y=6|X_1,\ldots,X_n) &= \frac{P(X_1,\ldots,X_n|Y=6)P(Y=6)}{P(X_1,\ldots,X_n|Y=5)P(Y=5)+P(X_1,\ldots,X_n|Y=6)P(Y=6)}
\end{aligned}
$$

- To classify, we’ll simply compute these two probabilities and predict based on which one is greater

# Model Parameters

- The problem with explicitly modeling $P(X_1,\ldots,X_n|Y)$ is that there are usually way too many parameters:
  - We'll run out of space
  - We'll run out of time
  - And we'll need tons of training data (which is usually not available)

# The Naïve Bayes Model

- The *Naïve Bayes Assumption*: Assume that all features are independent **given the class label Y**
- Equationally speaking:

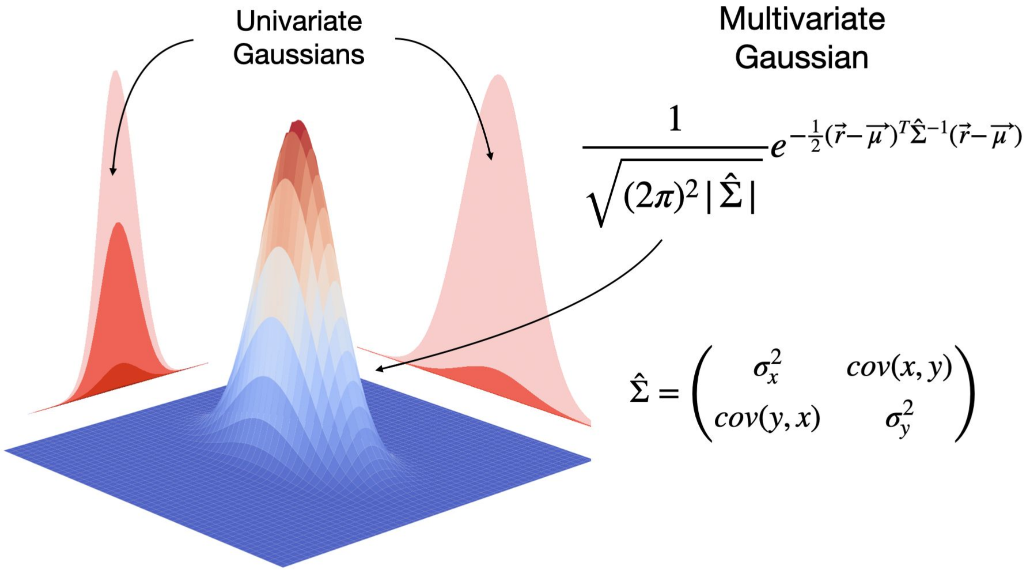
$$P(X_1, \ldots, X_n|Y) = \prod_{i=1}^{n} P(X_i|Y)$$

- (We will discuss the validity of this assumption later)

# The Naïve Bayes Model

- The *Naïve Bayes Assumption*: Assume that all features are independent **given the class label Y**
- Equationally speaking:

$$P(X_1,\ldots,X_n|Y)=\prod_{i=1}^{n}P(X_i|Y)$$

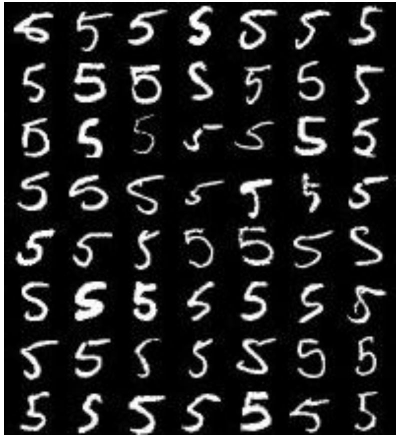


$$\frac{1}{\sqrt{(2\pi)^2|\hat{\Sigma}|}}e^{-\frac{1}{2}(\vec{r}-\vec{\mu})^T\hat{\Sigma}^{-1}(\vec{r}-\vec{\mu})}$$

$$\hat{\Sigma}=\begin{pmatrix}\sigma_x^2 & cov(x,y)\\ cov(y,x) & \sigma_y^2\end{pmatrix}$$

# Naïve Bayes Training

- Now that we've decided to use a Naïve Bayes classifier, we need to train it with some data:



MNIST Training Data

# Why is this useful?

- # of parameters for modeling $P(X_1,\ldots,X_n|Y)$:
  - $2(2^n-1)$
- # of parameters for modeling $P(X_1|Y),\ldots,P(X_n|Y)$
  - $2n$

# Naïve Bayes Training

- Training in Naïve Bayes is **easy**:
  - Estimate P(Y=v) as the fraction of records with Y=v

$$P(Y=v) = \frac{Count(Y=v)}{\#\ records}$$

  - Estimate P($X_i$=u|Y=v) as the fraction of records with Y=v for which $X_i$=u

$$P(X_i=u|Y=v) = \frac{Count(X_i=u \wedge Y=v)}{Count(Y=v)}$$

- (This corresponds to Maximum Likelihood estimation of model parameters)

# Naïve Bayes Training

- In practice, some of these counts can be zero
- Fix this by adding “virtual” counts (for two class problem):

$$P(X_i = u|Y = v) = \frac{Count(X_i = u \wedge Y = v) + 1}{Count(Y = v) + 2}$$

  - This is called *Smoothing*

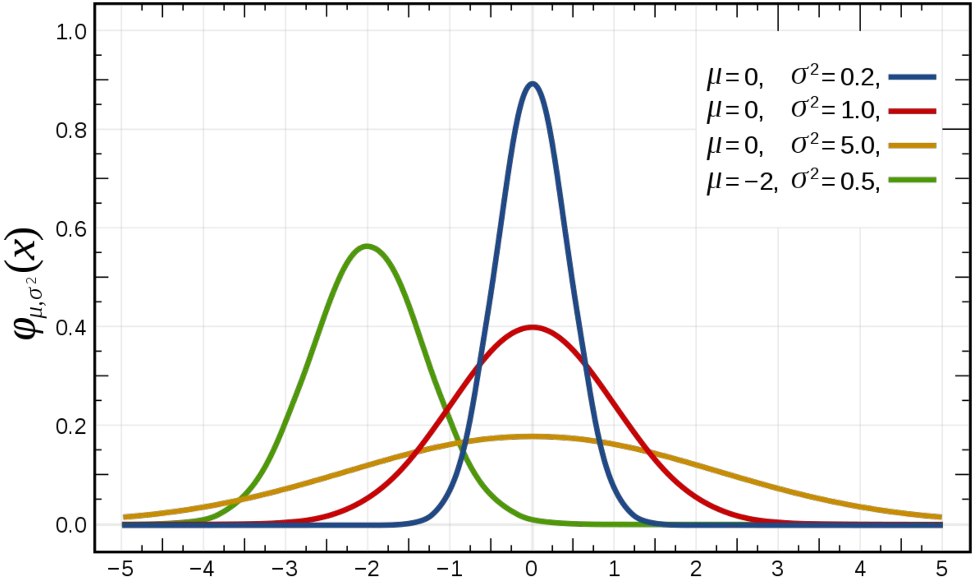
# The Naive Bayes Classifier for Data Sets with Numerical Attribute Values

- One common practice to handle numerical attribute values is to assume normal distributions for numerical attributes.
- Let $x_1, x_2, \ldots, x_n$ be the values of a numerical attribute in the training data set.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\sigma = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \mu\right)^2$$

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



- For examples,

$$f\left(\text{temperature} = 66 \mid \text{Yes}\right) = \frac{1}{\sqrt{2\pi}\left(6.2\right)} e^{-\frac{(66-73)^2}{2(6.2)^2}} = 0.0340$$

- Likelihood of Yes = $\frac{2}{9} \times 0.0340 \times 0.0221 \times \frac{3}{9} \times \frac{9}{14} = 0.000036$

- Likelihood of No = $\frac{3}{5} \times 0.0291 \times 0.038 \times \frac{3}{5} \times \frac{5}{14} = 0.000136$

# Numerical Stability

- It is often the case that machine learning algorithms need to work with very small numbers
  - Imagine computing the probability of 2000 independent coin flips
  - Python thinks that $(.5)^{2000}$=0

# Underflow Prevention

- Multiplying lots of probabilities

□ floating-point underflow.

- Recall: *log(xy) = log(x) + log(y),*

□ better to sum logs of probabilities rather than multiplying probabilities.

# Underflow Prevention

- Class with highest final un-normalized log probability score is still the most probable.

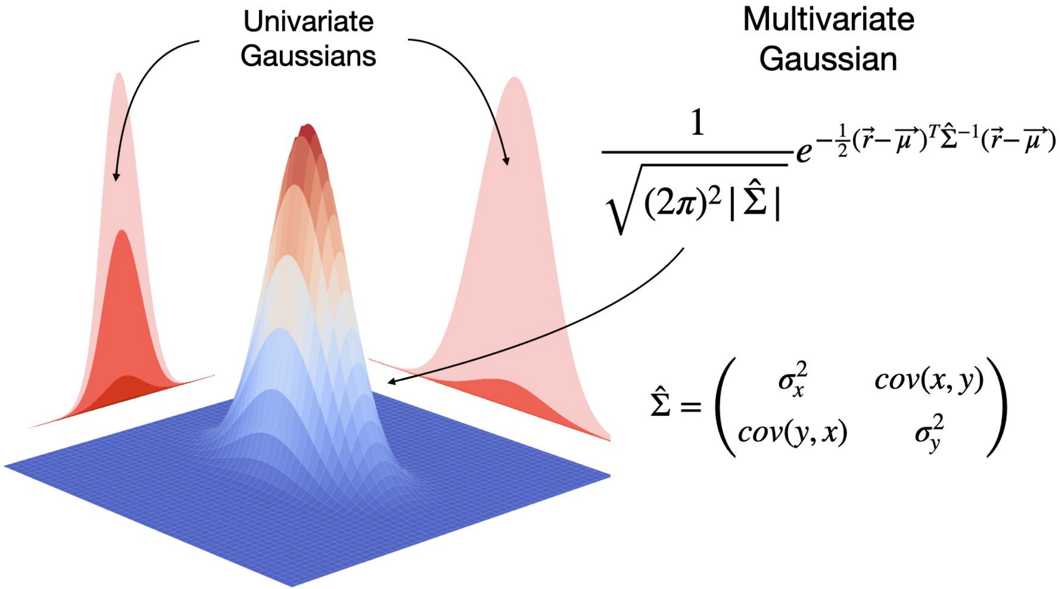
$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} \log P(c_j) + \sum_{i \in positions} \log P(x_i \mid c_j)$$

# Underflow Prevention

- Class with highest final un-normalized log probability score is still the most probable.

$$
\begin{aligned}
\log\left(P(Y|X_1,\ldots,X_n)\right) &= \log\left(\frac{P(X_1,\ldots,X_n|Y)\cdot P(Y)}{P(X_1,\ldots,X_n)}\right) \\
&= \text{constant} + \log\left(\prod_{i=1}^{n} P(X_i|Y)\right) + \log P( \\
&= \text{constant} + \sum_{i=1}^{n} \log P(X_i|Y) + \log P(Y)
\end{aligned}
$$

# Drawbacks



$$\frac{1}{\sqrt{(2\pi)^2|\hat{\Sigma}|}} e^{-\frac{1}{2}(\vec{r}-\vec{\mu})^T\hat{\Sigma}^{-1}(\vec{r}-\vec{\mu})}$$

$$\hat{\Sigma} = \begin{pmatrix} \sigma_x^2 & cov(x, y) \\ cov(y, x) & \sigma_y^2 \end{pmatrix}$$

- Actually, the Naïve Bayes assumption is almost never true
- Still… Naïve Bayes often performs surprisingly well even when its assumptions do not hold

# Merits

- Naïve Bayes is:
  - Really easy to implement and often works well
  - Often a good first thing to try
  - Commonly used as a “punching bag” for smarter algorithms