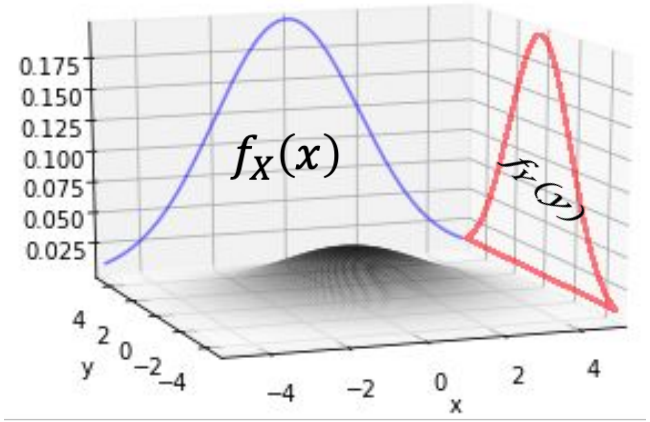
CS 109 Quiz 3 Review

If two random variables X and Y are jointly continuous, then there exists a joint probability density function $f_{X,Y}$ defined over $-\infty < x, y < \infty$ such that:

$$P(a_1 \leq X \leq a_2,\ \ b_1 \leq Y \leq b_2) = \int_{a_1}^{a_2} \int_{b_1}^{b_2} f_{X,Y}(x,y)\,dy\,dx$$

Suppose X and Y are continuous random variables with joint PDF:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dy\,dx = 1$$



The marginal density functions (marginal PDFs) are therefore:

$$f_X(a) = \int_{-\infty}^{\infty} f_{X,Y}(a,y)dy \qquad f_Y(b) = \int_{-\infty}^{\infty} f_{X,Y}(x,b)dx$$

For a continuous random variable X with PDF f, the CDF (cumulative distribution function) is

$$F(a) = P(X \leq a) = \int_{-\infty}^{a} f(x)dx$$

The density f is therefore the derivative of the CDF, F:

$$f(a) = \frac{d}{da}F(a)$$

(Fundamental Theorem of Calculus)

For two random variables X and Y, there can be a **joint cumulative distribution function** $F_{X,Y}$:

$$F_{X,Y}(a,b) = P(X \leq a, Y \leq b)$$

For discrete X and Y:

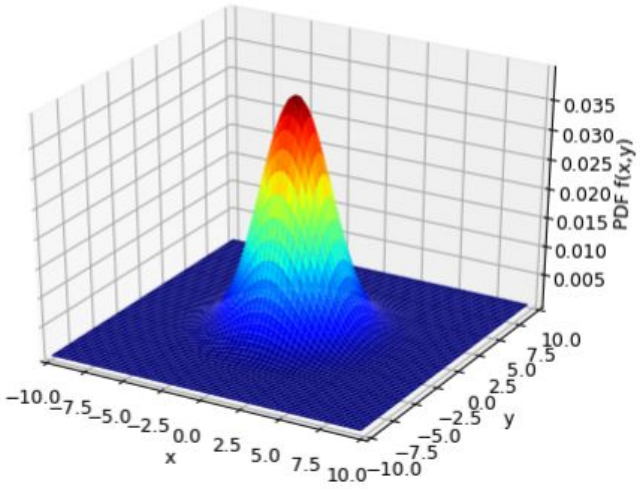
$$F_{X,Y}(a,b) = \sum_{x \leq a} \sum_{y \leq b} p_{X,Y}(x,y)$$

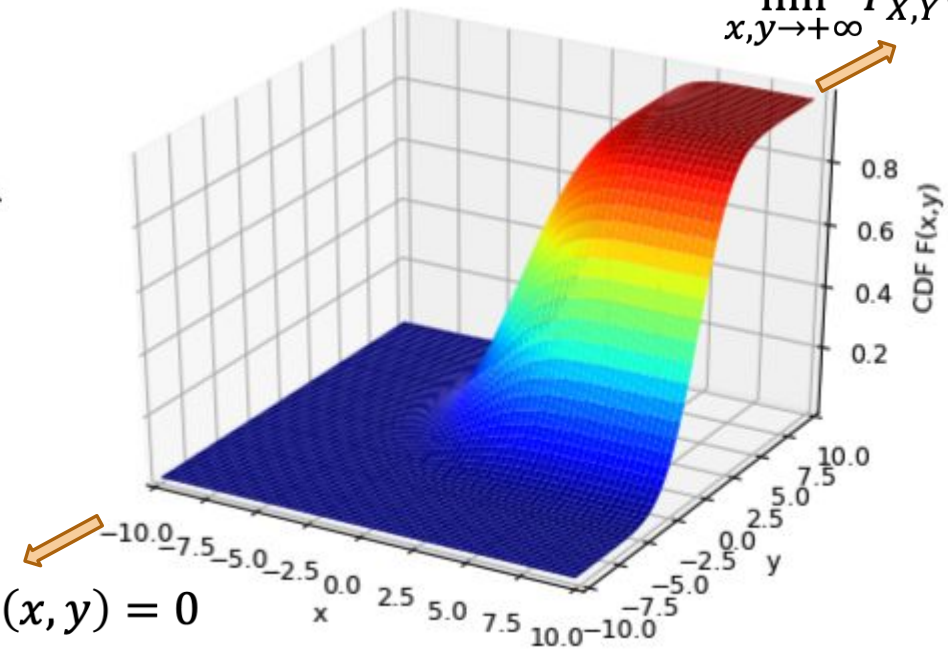
For continuous X and Y:

$$F_{X,Y}(a,b) = \int_{-\infty}^{a} \int_{-\infty}^{b} f_{X,Y}(x,y)\,dy\,dx$$

$$f_{X,Y}(a,b) = \frac{\partial^2}{\partial a\,\partial b} F_{X,Y}(a,b)$$

$$\lim_{x,y\to+\infty} F_{X,Y}(x,y) = 1$$





$$\lim_{x,y\to-\infty} F_{X,Y}(x,y) = 0$$

$$f_{X,Y}(x,y)$$

$$F_{X,Y}(x,y) = P(X \le x, Y \le y)$$

Recall for a single RV X with CDF F_X:

CDF: $P(X \leq x) = F_X(x)$

$$P(a < X \leq b) = F_X(b) - F(a)$$

For two RVs X and Y with joint CDF $F_{X,Y}$:

Joint CDF: $P(X \leq x, Y \leq y) = F_{X,Y}(x, y)$

$$P(a_1 < X \leq a_2, b_1 < Y \leq b_2) = F_{X,Y}(a_2, b_2) - F_{X,Y}(a_1, b_2) - F_{X,Y}(a_2, b_1) + F_{X,Y}(a_1, b_1)$$

Note strict inequalities; these properties hold for both discrete and continuous RVs.

Two continuous random variables X and Y are independent if:

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$$

Equivalently:

$$F_{X,Y}(x, y) = F_X(x)F_Y(y)$$
$$f_{X,Y}(x, y) = f_X(x)f_Y(y)$$

More generally, X and Y are independent if joint density factors separately:

$$f_{X,Y}(x, y) = g(x)h(y), \text{ where } -\infty < x, y < \infty$$

X_1 and X_2 follow a bivariate normal distribution if their joint PDF f is

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right)}$$

Can show that $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2), X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$

Often written as:

$$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- Vector $\boldsymbol{X} = (X_1, X_2)$
- Mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2)$, Covariance matrix: $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$

Recall correlation: $\rho = \frac{\text{Cov}(X_1, X_2)}{\sigma_1\sigma_2}$

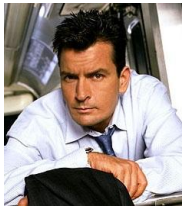
We will focus on understanding the **shape** of a bivariate Normal RV.

with events

$$P(F|E) = \frac{P(F)P(E|F)}{P(E)}$$

with discrete RVs

$$p_{Y|X}(y|x) = \frac{p_Y(y)p_{X|Y}(x|y)}{p_X(x)}$$



with continuous RVs

$$f_{Y|X}(y|x) = \frac{f_Y(y)f_{X|Y}(x|y)}{f_X(x)}$$

Really all the same idea!

M,N are discrete. X, Y are continuous

OG Bayes

$$p_{M|N}(m|n) = \frac{P_{N|M}(n|m)p_M(m)}{p_N(n)}$$

Mix Bayes #1

$$f_{X|N}(x|n) = \frac{P_{N|X}(n|x)f_X(x)}{P_N(n)}$$

Mix Bayes #2

$$p_{N|X}(n|x) = \frac{f_{X|N}(x|n)p_N(n)}{f_X(x)}$$

All Continuous

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)}$$

Recall that two events A and B are conditionally independent given E if:

$$P(AB|E) = P(A|E)P(B|E)$$

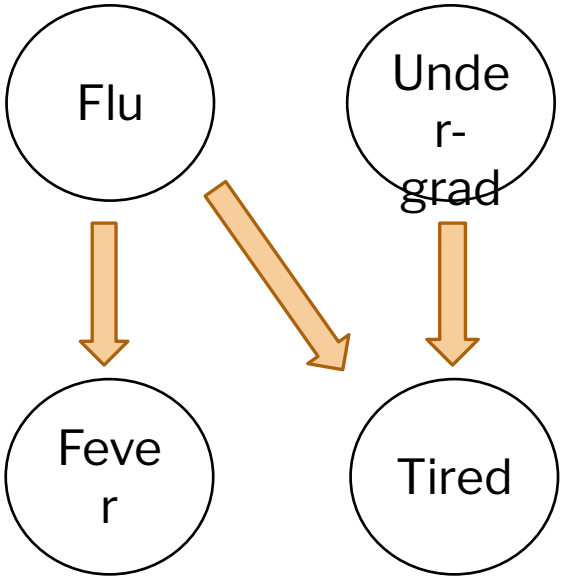
n discrete random variables $X_1, X_2, \ldots, X_n$ are called **conditionally independent given Y** if:

for all $x_1, x_2, \ldots, x_n, y$:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n | Y = y) = \prod_{i=1}^{n} P(X_i = x_i | Y = y)$$

This implies the following (cool to remember for later):

$$\log P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n | Y = y) = \sum_{i=1}^{n} \log P(X_i = x_i | Y = y)$$



In a Bayesian Network,

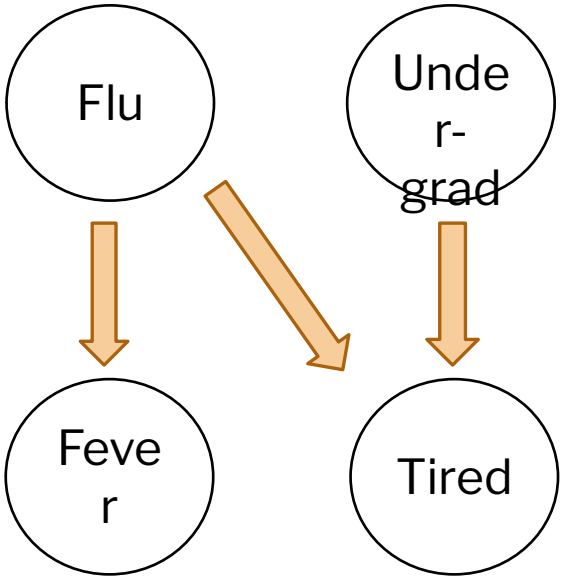
Each random variable is conditionally independent of its non-descendants, given its parents.

- Node: random variable
- Directed edge: conditional dependency

Examples:

- $P(F_{ev} = 1 | T = 0, F_{lu} = 1) = P(F_{ev} = 1 | F_{lu} = 1)$
- $P(F_{lu} = 1, U = 0) = P(F_{lu} = 1)P(U = 0)$

$P(F_{lu} = 1) = 0.1$ $P(U = 1) = 0.8$



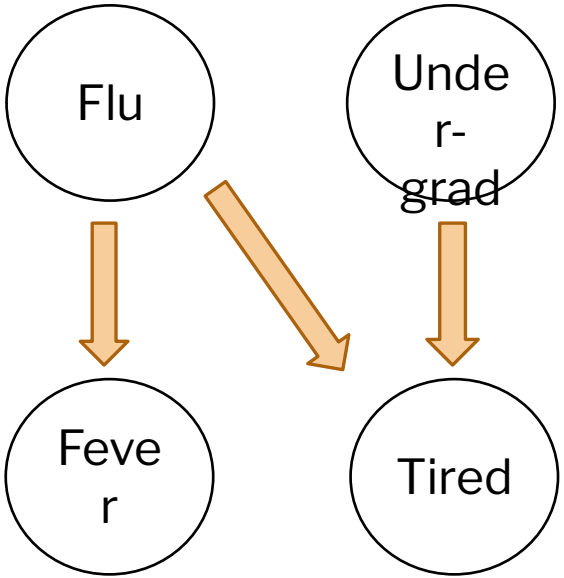
1. $P(F_{lu} = 0, U = 1, F_{ev} = 0, T = 1)$?

Compute joint probabilities using chain rule.

$P(F_{ev} = 1|F_{lu} = 1) = 0.9$
$P(F_{ev} = 1|F_{lu} = 0) = 0.05$

$P(T = 1|F_{lu} = 0, U = 0) = 0.1$
$P(T = 1|F_{lu} = 0, U = 1) = 0.8$
$P(T = 1|F_{lu} = 1, U = 0) = 0.9$
$P(T = 1|F_{lu} = 1, U = 1) = 1.0$

$P(F_{lu} = 1) = 0.1$ $P(U = 1) = 0.8$



2. $P(F_{lu} = 1 | F_{ev} = 0, U = 0, T = 1)$?

1. Compute joint probabilities

$P(F_{lu} = 1, F_{ev} = 0, U = 0, T = 1)$

$P(F_{lu} = 0, F_{ev} = 0, U = 0, T = 1)$

2. Definition of conditional probability

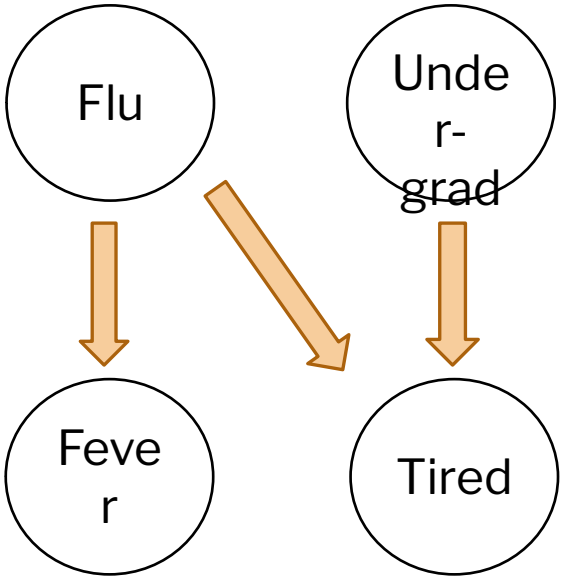
$$\frac{P(F_{lu} = 1, F_{ev} = 0, U = 0, T = 1)}{\sum_x P(F_{lu} = x, F_{ev} = 0, U = 0, T = 1)}$$

$P(F_{ev} = 1 | F_{lu} = 1) = 0.9$
$P(F_{ev} = 1 | F_{lu} = 0) = 0.05$

$P(T = 1 | F_{lu} = 0, U = 0) = 0.1$
$P(T = 1 | F_{lu} = 0, U = 1) = 0.8$
$P(T = 1 | F_{lu} = 1, U = 0) = 0.9$
$P(T = 1 | F_{lu} = 1, U = 1) = 1.0$

$= 0.095$

$P(F_{lu} = 1) = 0.1$ $P(U = 1) = 0.8$



3. $P(F_{lu} = 1 | U = 1, T = 1)$?

1. Compute joint probabilities

$P(F_{lu} = 1, U = 1, F_{ev} = 1, T = 1)$

...

$P(F_{lu} = 0, U = 1, F_{ev} = 0, T = 1)$?

2. Definition of conditional probability

$$\frac{\sum_y P(F_{lu} = 1, U = 1, F_{ev} = y, T = 1)}{\sum_x \sum_y P(F_{lu} = x, U = 1, F_{ev} = y, T = 1)}$$

$P(F_{ev} = 1 | F_{lu} = 1) = 0.9$
$P(F_{ev} = 1 | F_{lu} = 0) = 0.05$

$P(T = 1 | F_{lu} = 0, U = 0) = 0.1$
$P(T = 1 | F_{lu} = 0, U = 1) = 0.8$
$P(T = 1 | F_{lu} = 1, U = 0) = 0.9$
$P(T = 1 | F_{lu} = 1, U = 1) = 1.0$

$= 0.122$

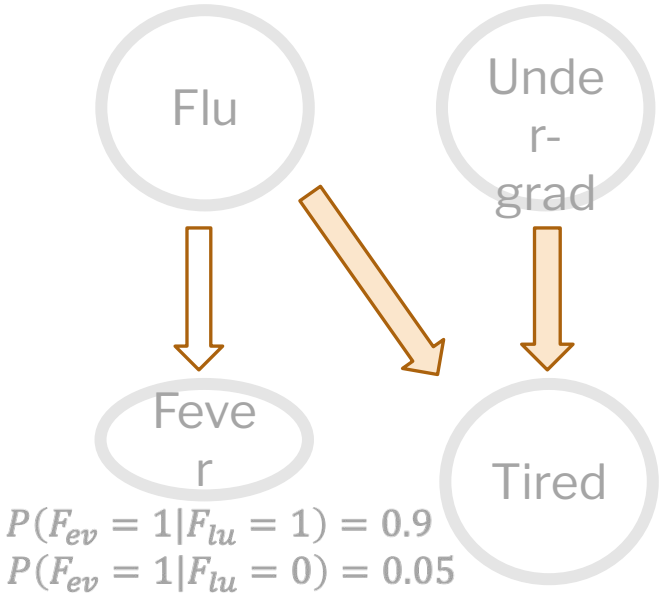
```
# Method: Make Sample
# ------------------
# create a single sample from the joint distribution
# based on the medical "WebMD" Bayesian Network
def make_sample():
    # prior on causal factors
    flu = bernoulli(0.1)
    und = bernoulli(0.8)

    # choose fever based on flu
    if flu == 1: fev = bernoulli(0.9)
    else:        fev = bernoulli(0.05)

    # choose tired based on (undergrad and flu)
    if   flu == 0 and und == 0: tir = bernoulli(0.1)
    elif flu == 0 and und == 1: tir = bernoulli(0.8)
    elif flu == 1 and und == 0: tir = bernoulli(0.9)
    else:                       tir = bernoulli(1.0)

    # a sample from the joint has an
    # assignment to *all* random variables
    return [flu, und, fev, tir]
```

$P(F_{lu} = 1) = 0.1$ $P(U = 1) = 0.8$



$P(F_{ev} = 1|F_{lu} = 1) = 0.9$
$P(F_{ev} = 1|F_{lu} = 0) = 0.05$

$P(T = 1|F_{lu} = 0, U = 0) = 0.1$
$P(T = 1|F_{lu} = 0, U = 1) = 0.8$
$P(T = 1|F_{lu} = 1, U = 0) = 0.9$
$P(T = 1|F_{lu} = 1, U = 1) = 1.0$

Are $X_1, X_2, \ldots, X_n$ i.i.d. with the following distributions?

1. $X_i \sim \text{Exp}(\lambda)$, X_i independent ✓
2. $X_i \sim \text{Exp}(\lambda_i)$, X_i independent ✗ (unless λ_i equal)
3. $X_i \sim \text{Exp}(\lambda)$, $X_1 = X_2 = \cdots = X_n$ ✗ dependent: $X_1 = X_2 = \cdots = X_n$
4. $X_i \sim \text{Bin}(n_i, p)$, X_i independent ✗ (unless n_i equal)
   Note underlying Bernoulli RVs are i.i.d.!

Consider n independent and identically distributed (i.i.d) variables $X_1, X_2, \ldots, X_n$ with $E[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$.

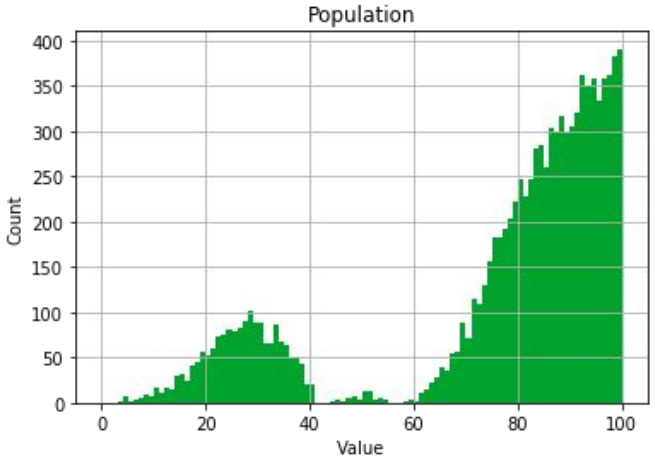
As $n \to \infty$

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$

The sum of n i.i.d. random variables is normally distributed with mean $n\mu$ and variance $n\sigma^2$.

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2) \quad \text{As } n \to \infty$$

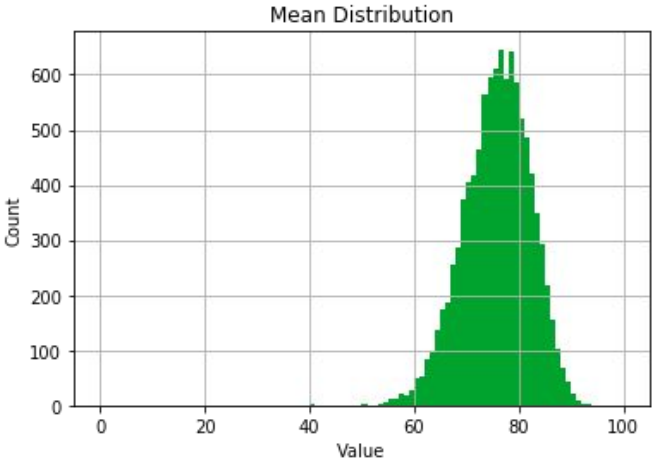
The sum of n **i.i.d.** random variables is normally distributed with mean $n\mu$ and variance $n\sigma^2$.



Distribution of X_i

Sample of size 15, **average** values

(sample mean)



Distribution of $\frac{1}{15}\sum_{i=1}^{15} X_i$

Let $X_1, X_2, \ldots, X_n$ be i.i.d., where $E[X_i] = \mu, \text{Var}(X_i) = \sigma^2$. As $n \to \infty$:

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$

Sum of i.i.d. RVs

$$\frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

Average of i.i.d. RVs (sample mean)

Gumbel

(see Fisher-Tippett Gnedenko Theorem)

Max of i.i.d. RVs

Consider *n* random variables X_1, X_2, ... X_n

- X_i are all independently and identically distributed (I.I.D.)
- Have same distribution function *F* and $E[X_i] = \mu$
- We call sequence of X_i a **sample** from distribution *F*
- *How would you estimate the population mean??*

$$\text{Estimate} = \frac{1}{n}\sum_{i=0}^{n} X_i$$

Sample Mean: This is a fancy way of saying "your estimate of the mean"

$$\bar{X} = \frac{1}{n}\sum_{i=0}^{n} X_i$$

If we knew the entire population $(x_1, x_2, \ldots, x_N)$:

population mean

population variance

$$\sigma^2 = E[(X-\mu)^2] = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$$

If we only have a sample, $(X_1, X_2, \ldots, X_n)$:

sample mean

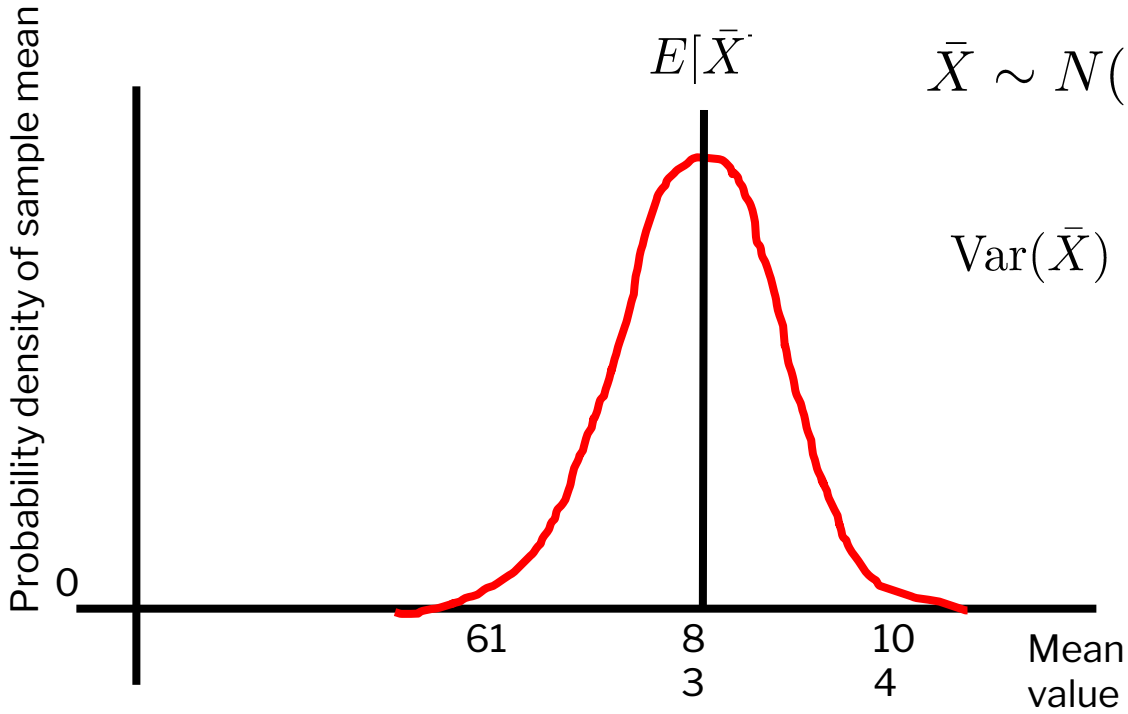
sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

By central limit theorem:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$



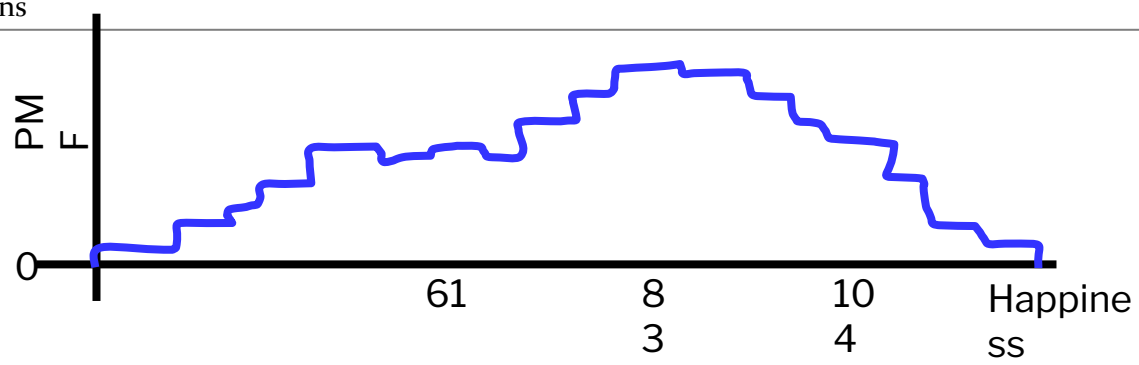
Bootstrapping Assumption

$$F \approx \hat{F}$$

The underlying distribution

The sample distribution

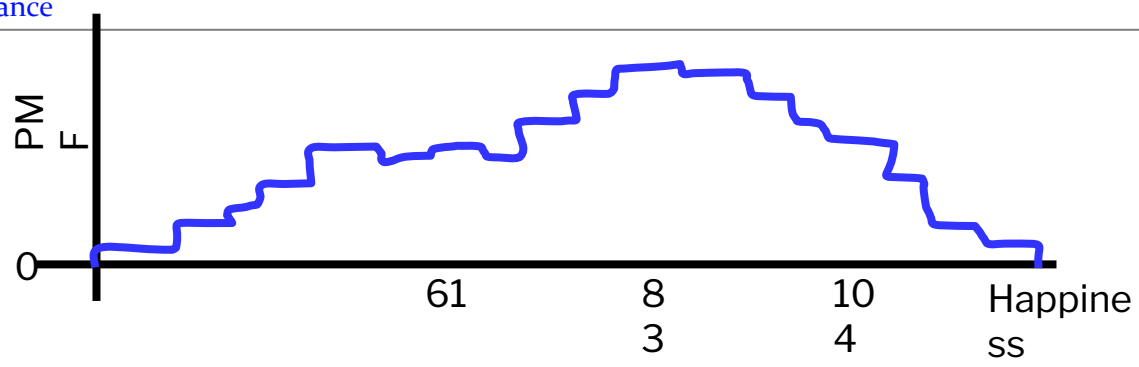
(aka the histogram of your data)



Bootstrap Algorithm (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
   a. Draw **sample.size()** new samples from PMF
   b. **Recalculate the mean on the resample**
3. You now have a **distribution of your means**

Means = [82.7, 83.4, 82.9, 91.4, 79.3, 82.1, …, 81.7]



Bootstrap Algorithm (sample):

1. Estimate the **PMF** using the sample
2. Repeat **10,000** times:
   a. Draw **sample.size()** new samples from PMF
   b. **Recalculate the var on the resample**
3. You now have a **distribution of your vars**

Vars = [472.7, 478.4, 469.2, …, 476.2]

Let X be continuous random variable

Let E be an event:

$$
\begin{aligned}
P(E|X=x) &= \frac{P(X=x,E)}{P(X=x)} \\
&= \frac{P(X=x|E)P(E)}{P(X=x)} \\
&= \frac{f_X(x|E)P(E)\epsilon_x}{f_X(x)\epsilon_x} \\
&= \frac{f_X(x|E)P(E)}{f_X(x)}
\end{aligned}
$$

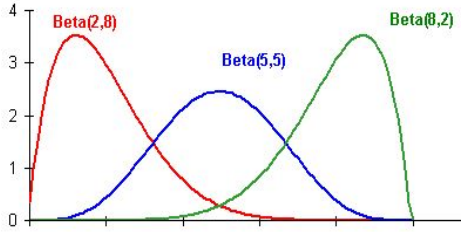
X is a **Beta Random Variable**: X ~ Beta(a, b)

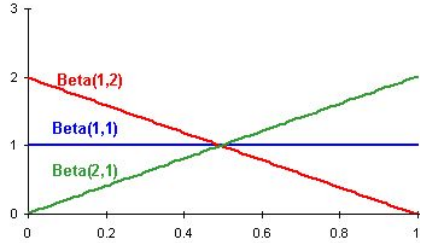
- Probability Density Function (PDF): (where $a, b > 0$)

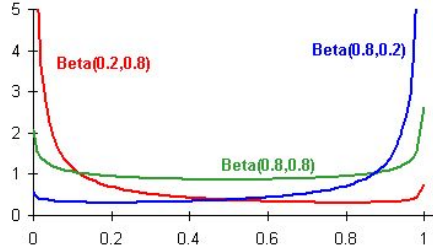
$$f(x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,dx$$

- Symmetric







$$E[X] = \frac{a}{a+b} \qquad Var(X) = \frac{ab}{(a+b)^2(a+b+1)}$$

A beta understanding

Can set X ~ Beta(a, b) as prior to reflect how biased you think coin is apriori

- This is a subjective probability (aka Bayesian)!
- Prior probability for X based on seeing (a + b – 2) "imaginary" trials, where
  (a – 1) of them were heads.
  (b – 1) of them were tails.
- Beta(1, 1) = Uni(0, 1) □ we haven't seen any "imaginary trials", so apriori know nothing about coin

Update to get posterior probability

- X | (n heads and m tails) ~ Beta(a + n, b + m)

def Many random variables we have learned so far are parametric models:

Distribution = model + parameter θ

ex The distribution Ber(0.2) = Bernoulli model, parameter $\theta = 0.2$.

For each of the distributions below, what is the parameter θ?

1. Ber(p) $\quad \theta = p$
2. Poi(λ) $\quad \theta = \lambda$
3. Uni(α, β) $\quad \theta = (\alpha, \beta)$
4. $\mathcal{N}(\mu, \sigma^2)$ $\quad \theta = (\mu, \sigma^2)$
5. $Y = mX + b$ $\quad \theta = (m, b)$

θ is the parameter of a distribution. Note that θ can be a vector.

Consider a sample of n i.i.d. random variables $X_1, X_2, \ldots, X_n$, drawn from a distribution $f(X_i|\theta)$.

def The **Maximum Likelihood Estimator (MLE)** of θ is the value of θ that maximizes $L(\theta)$.

$$\theta_{MLE} = \arg\max_{\theta} L(\theta)$$

Likelihood of your sample

$$L(\theta) = \prod_{i=1}^{n} f(X_i|\theta)$$

For continuous X_i, $f(X_i|\theta)$ is PDF; for discrete X_i, $f(X_i|\theta)$ is PMF

$$\arg\max_x f(x)$$

The argument x that maximizes the function $f(x)$.

$$= \arg\max_x \log f(x)$$

(log is an increasing function: $x < y \Leftrightarrow \log x < \log y$)

$$= \arg\max_x (c \log f(x))$$

$(x < y \Leftrightarrow c \log x < c \log y)$

for any positive constant c

$$\hat{x} = \arg\max_x f(x)$$

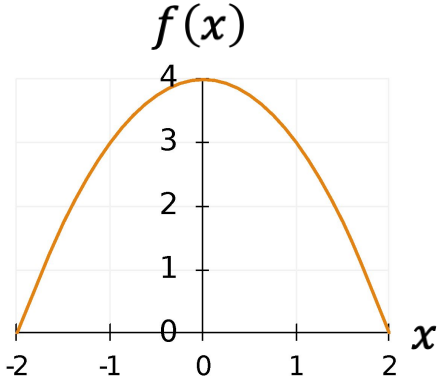
Let $f(x) = -x^2 + 4$, where $-2 < x < 2$.

Differentiate w.r.t. argmax's argument

$$\frac{d}{dx} f(x) = \frac{d}{dx}(x^2 + 4) = 2x$$

Set to 0 and solve

$$2x = 0 \quad \Rightarrow \quad \hat{x} = 0$$



Make sure $\hat{x}$ is a maximum

- Check $f(\hat{x} \pm \epsilon) < f(\hat{x})$
- Generally ignored in expository derivations
- We'll ignore it here too (and won't require it in class)
- arg min is defined similarly, relevant for gradient descent

Consider a sample of n i.i.d. random variables $X_1, X_2, \ldots, X_n$, drawn from a distribution $f(X_i|\theta)$.

$$L(\theta) = \prod_{i=1}^{n} f(X_i|\theta)$$

θ_{MLE} maximizes the likelihood of our sample, $L(\theta)$:

$$\theta_{MLE} = \arg\max_{\theta} L(\theta)$$

θ_{MLE} also maximizes the log-likelihood function, $LL(\theta)$:

$$\theta_{MLE} = \arg\max_{\theta} LL(\theta)$$

$$LL(\theta) = \log L(\theta) = \log\left(\prod_{i=1}^{n} f(X_i|\theta)\right) = \sum_{i=1}^{n} \log f(X_i|\theta)$$

$LL(\theta)$ is often easier to differentiate than $L(\theta)$.

Consider a sample of n i.i.d. RVs $X_1, X_2, \ldots, X_n$.

What is $\theta_{MLE} = p_{MLE}$?

- Let $X_i \sim \text{Ber}(p)$.
- $f(X_i|p) = p^{X_i}(1-p)^{1-X_i}$

1. Determine formula for $LL(\theta)$

$$LL(\theta) = \sum_{i=1}^{n} [X_i \log p + (1 - X_i) \log(1-p)]$$

$$= Y(\log p) + (n - Y) \log(1-p), \text{ where } Y = \sum_{i=1}^{n} X_i$$

2. Differentiate $LL(\theta)$ w.r.t. (each) θ, set to 0

$$\frac{\partial LL(\theta)}{\partial p} = Y\frac{1}{p} + (n - Y)\frac{-1}{1-p} = 0$$

3. Solve resulting equations

$$p_{MLE} = \frac{1}{n}Y = \frac{1}{n}\sum_{i=1}^{n} X_i$$

MLE of the Bernoulli parameter, p_{MLE}, is the unbiased estimate of the mean, $\bar{X}$ (sample mean)

Consider a sample of n i.i.d. RVs $X_1, X_2, \ldots, X_n$.

What is $\theta_{MLE} = \lambda_{MLE}$?

- Let $X_i \sim \text{Poi}(\lambda)$.
- PMF: $f(X_i|\lambda) = \frac{e^{-\lambda}\lambda^{X_i}}{X_i!}$

1. Determine formula for $LL(\theta)$

$$LL(\theta) = \sum_{i=1}^{n} \log\left(\frac{e^{-\lambda}\lambda^{X_i}}{X_i!}\right) = \sum_{i=1}^{n} (-\lambda \log e + X_i \log \lambda - \log X_i!)$$

$$= -n\lambda + \log(\lambda)\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \log(X_i!)$$

(using natural log, $\ln e = 1$)

2. Differentiate $LL(\theta)$ w.r.t. (each) θ, set to 0

$$\frac{\partial LL(\theta)}{\partial \lambda} = -n + \frac{1}{\lambda}\sum_{i=1}^{n} X_i = 0$$

3. Solve resulting equations

$$\lambda_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

MLE of the Poisson parameter, λ_{MLE}, is the unbiased estimate of the mean, $\bar{X}$ (sample mean)

Consider a sample of n i.i.d. random variables $X_1, X_2, \ldots, X_n$.

- Let $X_i \sim \mathcal{N}(\mu, \sigma^2)$.

$$f(X_i|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(X_i-\mu)^2/(2\sigma^2)}$$

What is $\theta_{MLE} = (\mu_{MLE}, \sigma^2_{MLE})$?

3. Solve resulting equations

Two equations, two unknowns:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu) = 0 \qquad -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(X_i - \mu)^2 = 0$$

First, solve for μ_{MLE}:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} X_i - \frac{1}{\sigma^2}\sum_{i=1}^{n}\mu = 0 \Rightarrow \sum_{i=1}^{n} X_i = n\mu \Rightarrow \mu_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

unbiased

Next, solve for σ_{MLE}:

$$\frac{1}{\sigma^3}\sum_{i=1}^{n}(X_i - \mu)^2 = \frac{n}{\sigma} \Rightarrow \sum_{i=1}^{n}(X_i - \mu)^2 = \sigma^2 n \Rightarrow \sigma^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_{MLE})^2$$

biased

Consider a sample of n i.i.d. random variables $X_1, X_2, \ldots, X_n$ (data).

Maximum Likelihood Estimator (MLE)

What is the parameter θ that **maximizes the likelihood** of our observed data $(X_1, X_2, \ldots, X_n)$?

$$L(\theta) = f(X_1, X_2, \ldots, X_n|\theta) = \prod_{i=1}^{n} f(X_i|\theta)$$

$$\theta_{MLE} = \arg\max_{\theta} f(X_1, X_2, \ldots, X_n|\theta)$$

likelihood of data

Maximum a Posteriori (MAP) Estimator

Given our observed data $(X_1, X_2, \ldots, X_n)$, what is the **most likely parameter** θ?

$$\theta_{MAP} = \arg\max_{\theta} f(\theta|X_1, X_2, \ldots, X_n)$$

posterior distribution of θ

Consider a sample of n i.i.d. random variables $X_1, X_2, \ldots, X_n$ (data).

def The **Maximum a Posteriori (MAP) Estimator** of θ is the value of θ that maximizes the posterior distribution of θ.

$$\theta_{MAP} = \arg\max_{\theta} f(\theta | X_1, X_2, \ldots, X_n)$$

Intuition with Bayes' Theorem:

$L(\theta)$, probability of data given parameter θ

After seeing data, posterior belief of θ

$$\underbrace{P(\theta|\text{data})}_{\text{posterior}} = \frac{\overbrace{P(\text{data}|\theta)}^{\text{likelihood}}\overbrace{P(\theta)}^{\text{prior}}}{P(\text{data})}$$

Before seeing data, prior belief of θ

Solving for θ_{MAP}

- Observe data: $X_1, X_2, \ldots, X_n$, all i.i.d.
- Let likelihood be same as MLE: $f(X_1, X_2, \ldots, X_n|\theta) = \prod_{i=1}^{n} f(X_i|\theta)$
- Let the prior distribution of θ be $g(\theta)$.

$$\theta_{MAP} = \arg\max_{\theta} f(\theta|X_1, X_2, \ldots, X_n) = \arg\max_{\theta} \frac{f(X_1, X_2, \ldots, X_n|\theta)g(\theta)}{h(X_1, X_2, \ldots, X_n)}$$ (Bayes' Theorem)

$$= \arg\max_{\theta} \frac{g(\theta)\prod_{i=1}^{n} f(X_i|\theta)}{h(X_1, X_2, \ldots, X_n)}$$ (independence)

$$= \arg\max_{\theta} g(\theta)\prod_{i=1}^{n} f(X_i|\theta)$$ ($1/h(X_1, X_2, \ldots, X_n)$ is a positive constant w.r.t. θ)

$$= \arg\max_{\theta} \left(\log g(\theta) + \sum_{i=1}^{n} \log f(X_i|\theta)\right)$$

Beta is a conjugate distribution for Bernoulli, meaning:

- Parametric forms of prior and posterior are the same
- Practically, conjugate means easy update:
  Add numbers of "successes" and "failures" seen to Beta parameters.
- You can set the prior to reflect how fair/biased you think the experiment is apriori.

Prior	$\text{Beta}(a = n_{imag} + 1, b = m_{imag} + 1)$
Experiment	Observe n successes and m failures
Posterior	$\text{Beta}(a = n_{imag} + n + 1, b = m_{imag} + m + 1)$

Mode of $\text{Beta}(a, b)$: $\dfrac{a-1}{a+b-2}$

(we'll prove this in a few minutes)

Beta parameters a, b are called **hyperparameters**.
Interpret $\text{Beta}(a, b)$: $a + b - 2$ trials,
of which $a - 1$ are successes

Observe data

Choose model with parameter θ

Choose prior on θ

Find $\theta_{MAP} = \arg\max_{\theta} f(\theta|X_1, X_2, \ldots, X_n)$

$$= \arg\max_{\theta} \left(\log g(\theta) + \sum_{i=1}^{n} \log f(X_i|\theta) \right)$$

Two valid approaches to computing θ_{MAP}

Mode of posterior distribution of θ

or

maximize log prior + log-likelihood

If we choose a conjugate prior, we avoid calculus with MAP: just report mode of posterior.

MAP with **Laplace smoothing**: a prior which represents k imagined observations of each outcome.

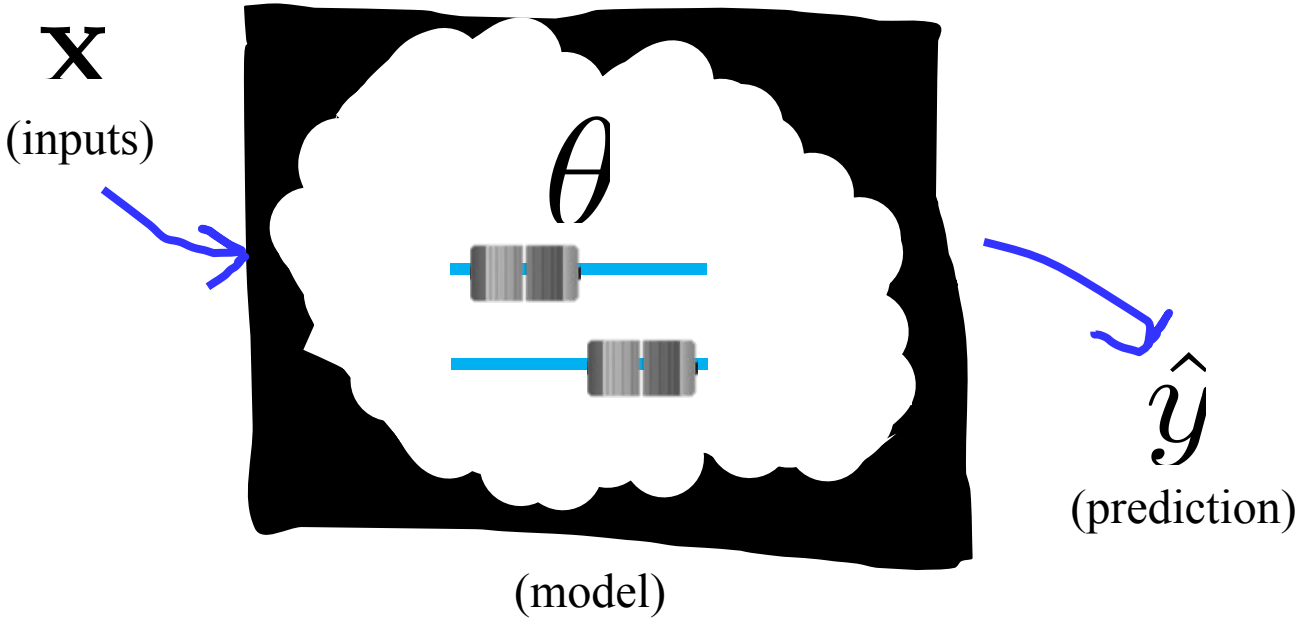
- Categorical data (i.e., Multinomial, Bernoulli/Binomial)
- Also known as additive smoothing

Laplace estimate Imagine $k = 1$ of each outcome
(follows from Laplace's "law of succession")

Example: Laplace estimate for coin probabilities from aforementioned experiment (100 coins: 58 heads, 42 tails)

heads $\frac{59}{102}$ tails $\frac{43}{102}$

Laplace smoothing:
- Easy to implement/remember


x
(inputs)
θ
ŷ
(prediction)
(model)

Training Data

Training Data: assignments all random variables **X** and Y

Assume IID data:

n training datapoints

$$(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \ldots (\mathbf{x}^{(n)}, y^{(n)})$$

$$m = |\mathbf{x}^{(i)}$$

Each datapoint has m features and a single output

Brute Force Bayes

$$\hat{y} = \underset{y=\{0,1\}}{\operatorname{argmax}} P(y|\mathbf{x})$$

$$= \underset{y=\{0,1\}}{\operatorname{argmax}} \frac{P(\mathbf{x}|y)P(y)}{P(\mathbf{x})}$$

$$= \underset{y=\{0,1\}}{\operatorname{argmax}} P(\mathbf{x}|y)P(y)$$

Simply chose the class label that is the most likely given the data

This is for one user

* Note how similar this is to Hamilton example 😊

Brute Force Bayes m = 100

Simply chose the class label that is the most likely given the data

$$\hat{y} = \underset{y=\{0,1\}}{\operatorname{argmax}} P(y|\mathbf{x})$$

$$= \underset{y=\{0,1\}}{\operatorname{argmax}} \frac{P(\mathbf{x}|y)P(y)}{P(\mathbf{x})}$$

$$= \underset{y=\{0,1\}}{\operatorname{argmax}} P(\mathbf{x}|y)P(y)$$

$$P(x_1, x_2, x_3, \ldots, x_{100}|y)$$

Big O of Brute Force Joint

What is the big O for # parameters?
m = # features.

$$O(2^m)$$

Assuming each feature is binary...

Naïve Bayes Assumption

$$\begin{aligned}\hat{y} &= \underset{y=\{0,1\}}{\operatorname{argmax}} \; P(y|\mathbf{x}) \\ &= \underset{y=\{0,1\}}{\operatorname{argmax}} \; \frac{P(\mathbf{x}|y)P(y)}{P(\mathbf{x})} \\ &= \underset{y=\{0,1\}}{\operatorname{argmax}} \; P(\mathbf{x}|y)P(y)\end{aligned}$$

$$\begin{aligned}P(\mathbf{x}|y) &= P(x_1, x_2, \ldots, x_m|y) \\ &= \prod_i P(x_i|y)\end{aligned}$$

The Naïve Bayes assumption

Naïve Bayes

Our prediction for y

Is a function of x

That choses the best value of y given x

$$\hat{y} = g(\mathbf{x}) = \underset{y \in \{0,1\}}{\operatorname{argmax}} \hat{P}(y|\mathbf{x})$$

$$= \underset{y \in \{0,1\}}{\operatorname{argmax}} \hat{P}(\mathbf{x}|y)\hat{P}(y)$$

Bayes rule!

$$= \underset{y}{\operatorname{argmax}} \left(\prod_{i=1}^{n} \hat{P}(x_i|y) \right) \hat{P}(y)$$

Naïve Bayes Assumption

$$= \underset{y}{\operatorname{argmax}} \log \hat{P}(y) + \sum_{i=1}^{m} \log \hat{P}(x_i|y)$$

This log version is useful for numerical stability

Computing Probabilities from Data

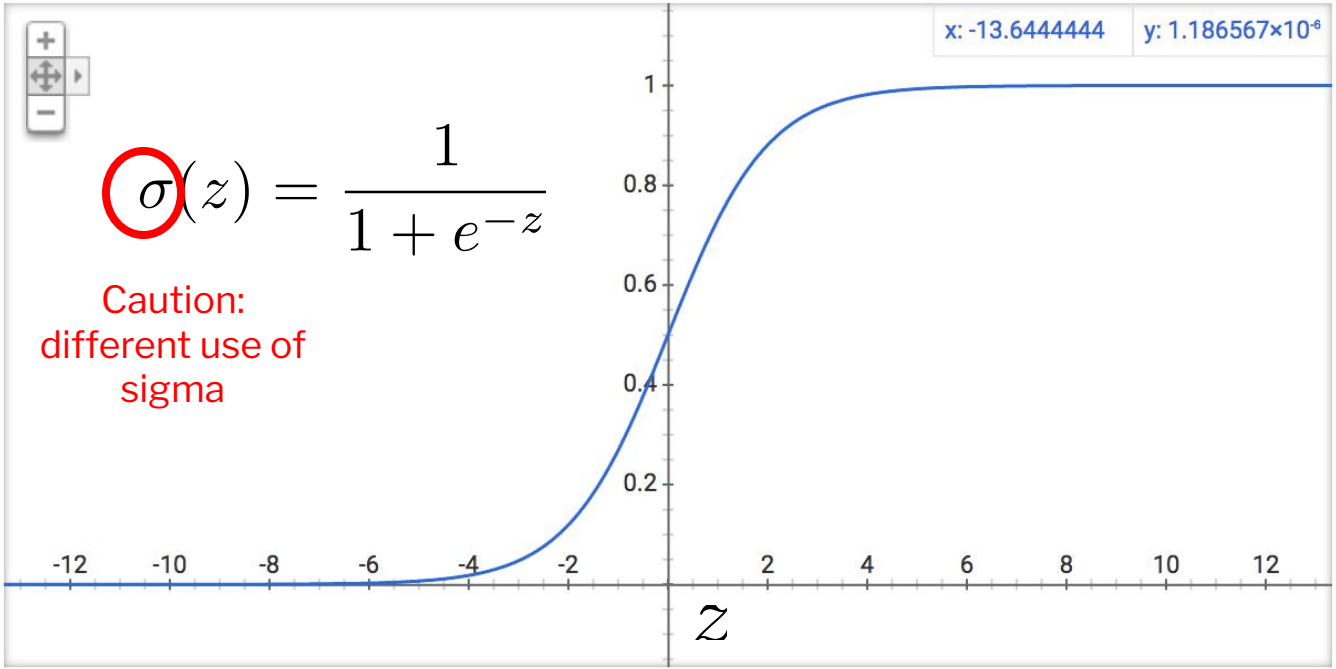
Various probabilities you will need to compute for Naive Bayesian Classifier (using MLE here):

$$\hat{p}(X_i = 1|Y = 0) = \frac{(\#\text{ training examples where } X_i = 1 \text{ and } Y = 0)}{(\#\text{ training examples where } Y = 0)}$$

$$\hat{p}(Y = 1) = \frac{(\#\text{ training examples where } Y = 1)}{(\#\text{ training examples})}$$

Logistic Regression

Background: Sigmoid Function

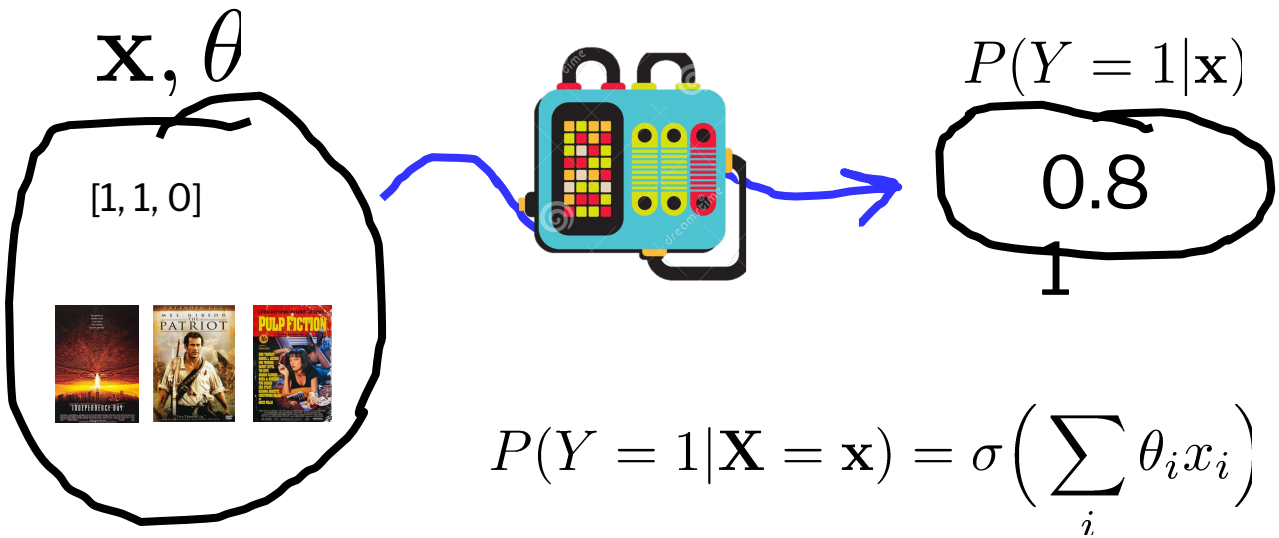


The sigmoid function squashes z to be a number between 0 and 1

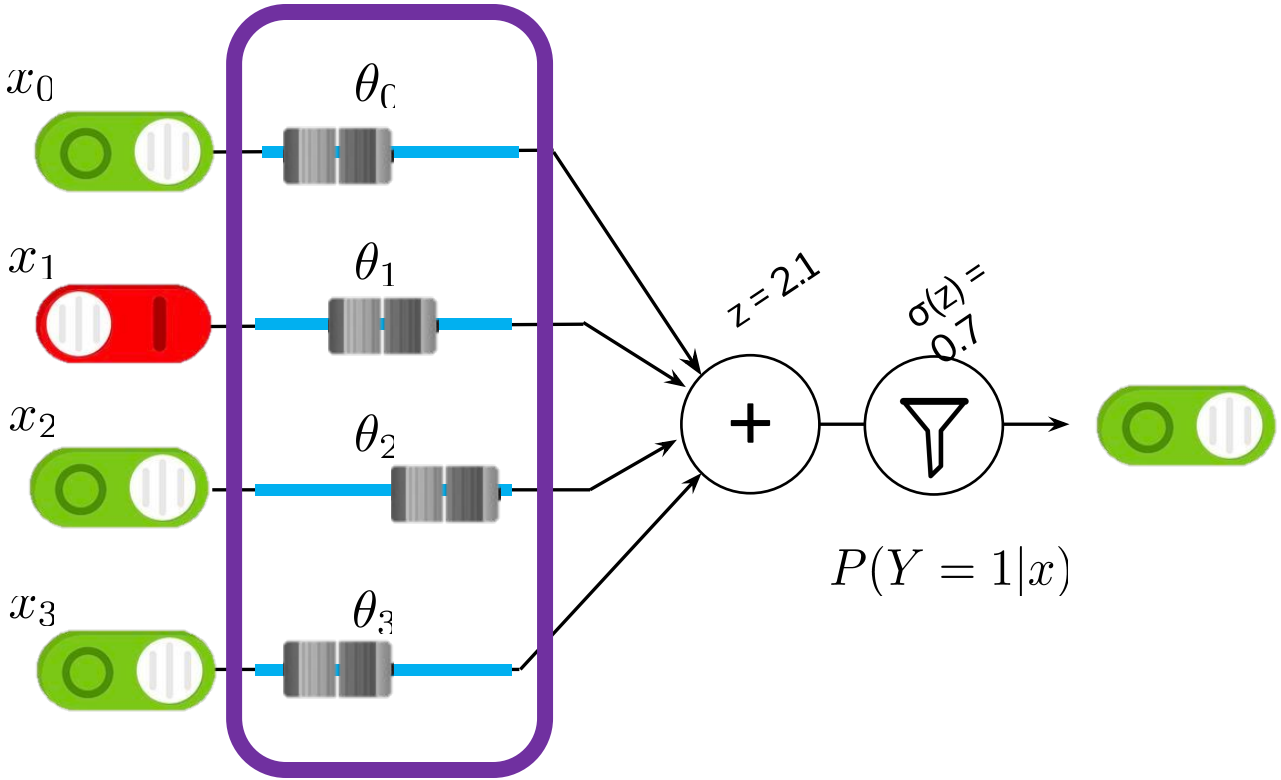
Logistic Regression Assumption

Could we model P(Y | **X**) directly?

- Welcome our friend: logistic regression!



Parameters Affect Prediction



$$P(Y = 1|\mathbf{X} = \mathbf{x}) = \sigma\left(\sum_i \theta_i x_i\right)$$

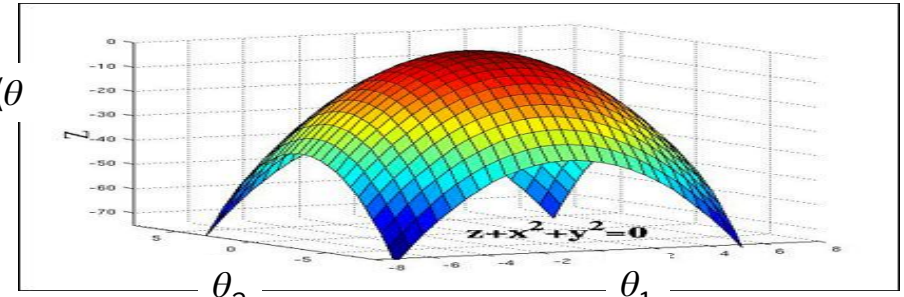
Gradient Ascent Step

$$\frac{\partial LL(\theta)}{\partial \theta_j} = \sum_{i=0}^{n} \left[y^{(i)} - \sigma(\theta^T \mathbf{x}^{(i)}) \right] x_j^{(i)}$$

$$\theta_j^{\text{new}} = \theta_j^{\text{old}} + \eta \cdot \frac{\partial LL(\theta^{\text{old}})}{\partial \theta_j^{\text{old}}}$$

$$= \theta_j^{\text{old}} + \eta \cdot \sum_{i=0}^{n} \left[y^{(i)} - \sigma(\theta^T \mathbf{x}^{(i)}) \right] x_j^{(i)}$$

Do this for all thet as!



Gradient Descent Step

Assume some loss function with known derivative $\frac{\partial \text{ Loss}}{\partial \theta_j}$

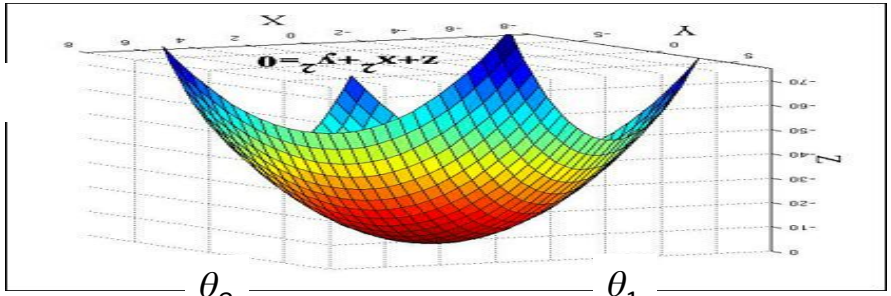
$$\theta_j^{\text{new}} = \theta_j^{\text{old}} - \eta \cdot \frac{\partial \text{ Loss}}{\partial \theta_j}$$

$$= \theta_j^{\text{old}} - \eta \cdot \frac{\partial \text{ Negative}LL}{\partial \theta_j}$$

$$= \theta_j^{\text{old}} + \eta \cdot \sum_{i=0}^{n} \left[y^{(i)} - \sigma(\theta^T \mathbf{x}^{(i)}) \right] x_j^{(i)}$$

...exactly the same

-LL(θ)



Logistic Regression Training

Initialize: $\theta_j = 0$ for all $0 \le j \le m$

Repeat many times:

gradient[j] = 0 for all $0 \le j \le m$

For each training example (x, y):

For each parameter j:

gradient[j] $+= x_j \left(y - \frac{1}{1 + e^{-\theta^T \mathbf{x}}} \right)$

θ_j += η * gradient[j] for all $0 \le j \le m$