08.03.2019

Statistical Methods in AI (CSE/ECE 471)

Lecture-17: MLE, MAP and Bayesian Estimation

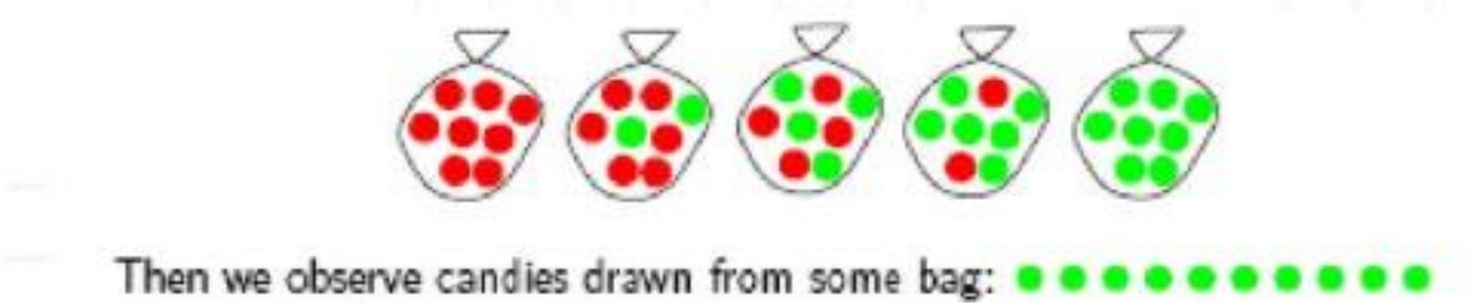
Ravi Kiran

Center for Visual Information Technology (CVIT), IIIT Hyderabad


PROBABILITY = EVENT / OUTCOMES
wikiHow to Calculate Probability

Data – a probability-based perspective

- The basis for Statistical Learning Theory

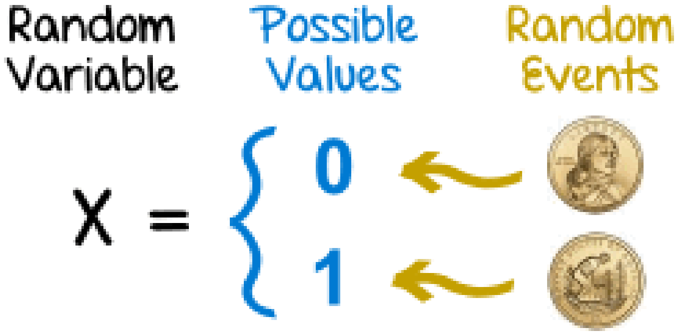


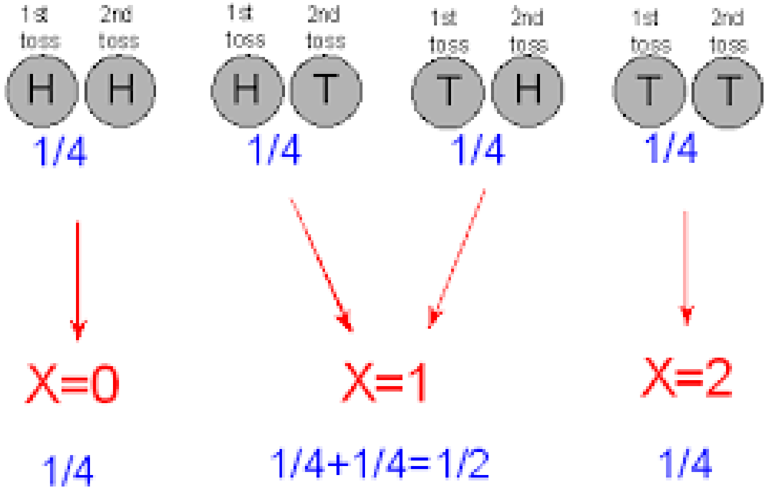
  – Domain described by random variables (r.v.)
    - X = {apple, grape}
    - $b_i \in [1,5]$

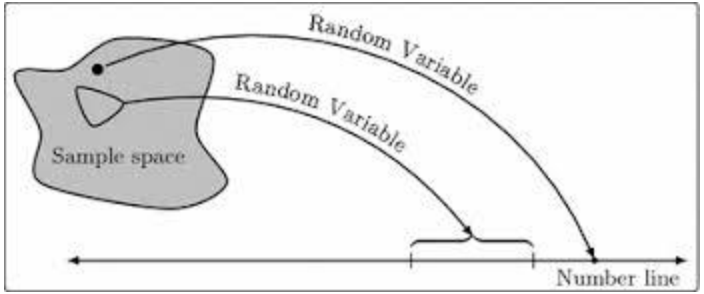
  – **Data = Instantiation of some or all r.v.'s in the domain**

Random Variables

R.V. = A numerical value from a random experiment



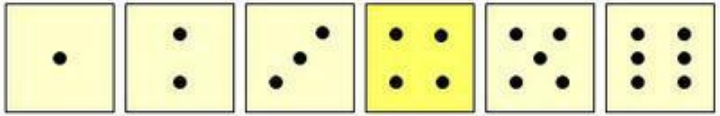




Discrete Random Variables

- Can only take on a countable number of values

Examples:

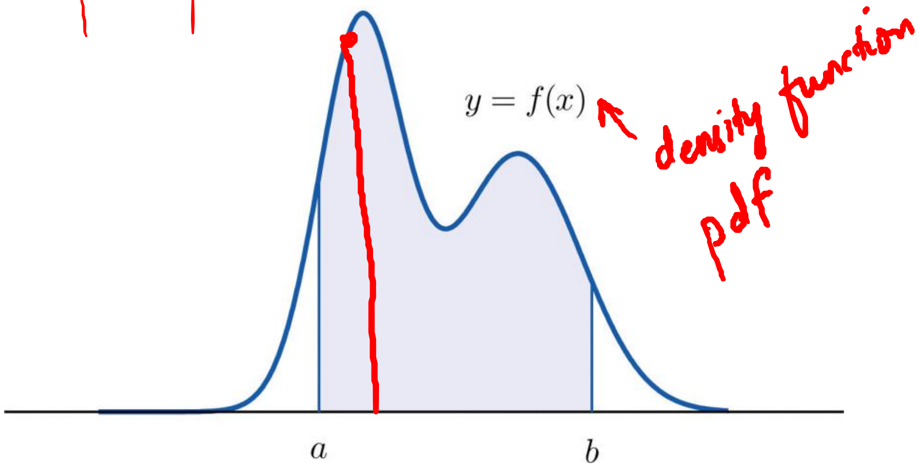


- **Roll a die twice**
  Let X be the number of times 4 comes up
  (then X could be 0, 1, or 2 times)

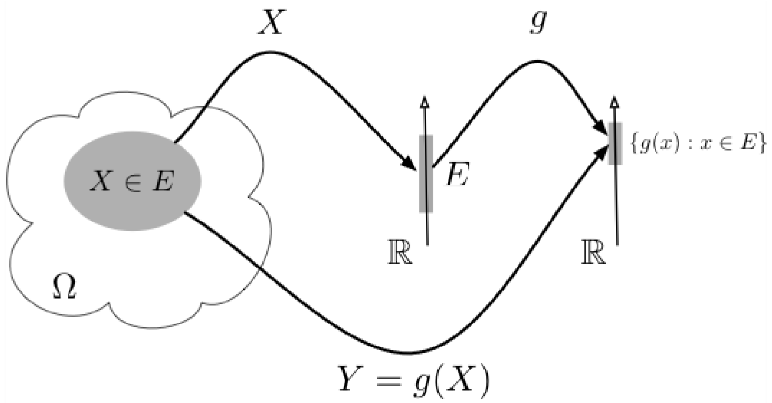
- **Toss a coin 5 times.**
  Let X be the number of heads
  (then X = 0, 1, 2, 3, 4, or 5)

Continuous random variable

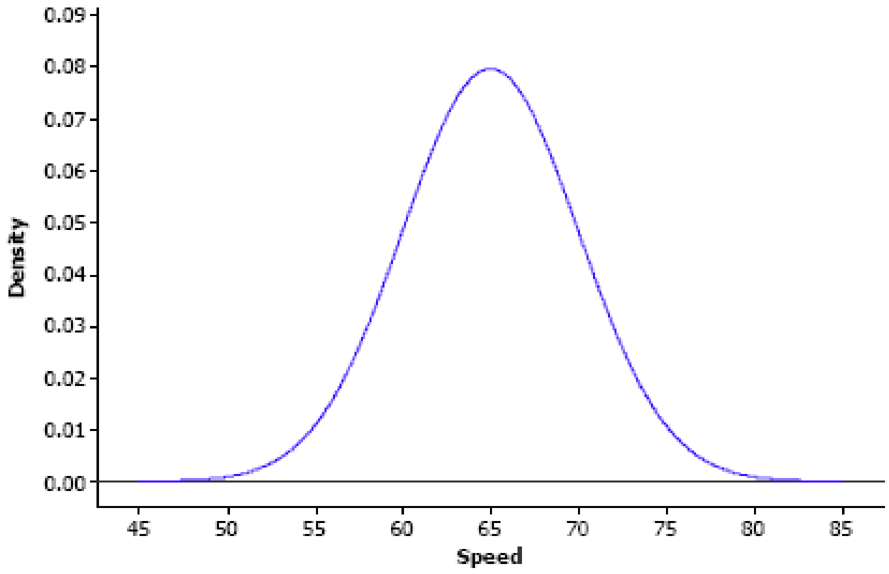
$P(a < X < b) =$ area of shaded region

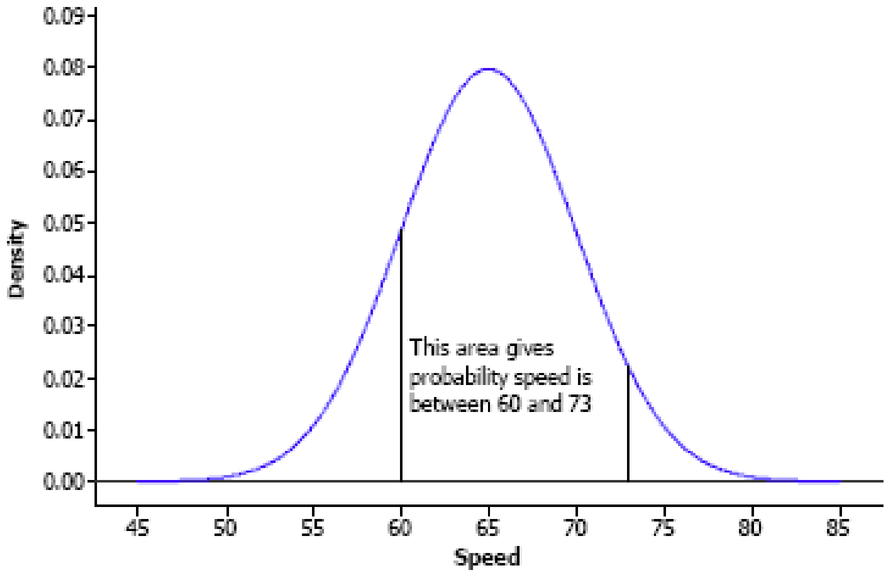


$P(Y = 0.2) = 0$

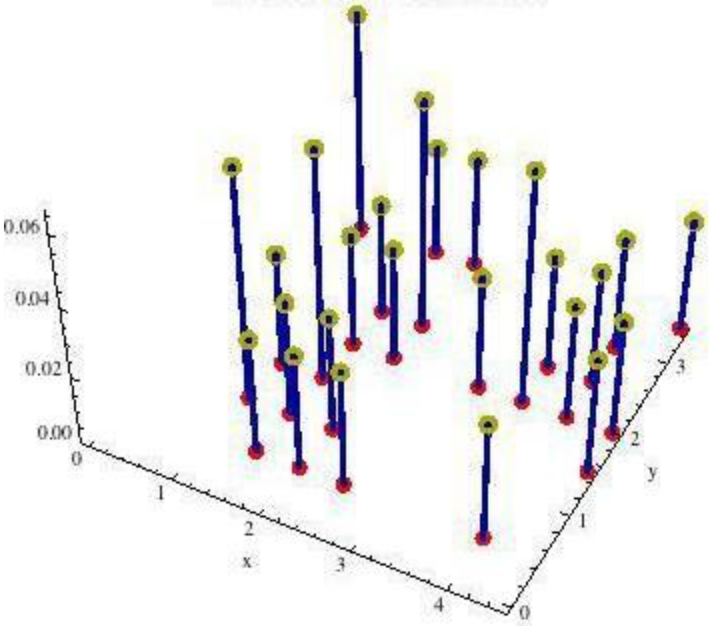


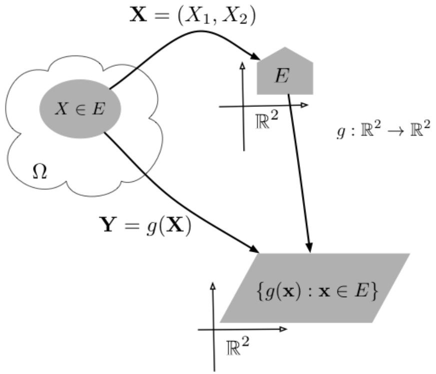
Continuous random variable

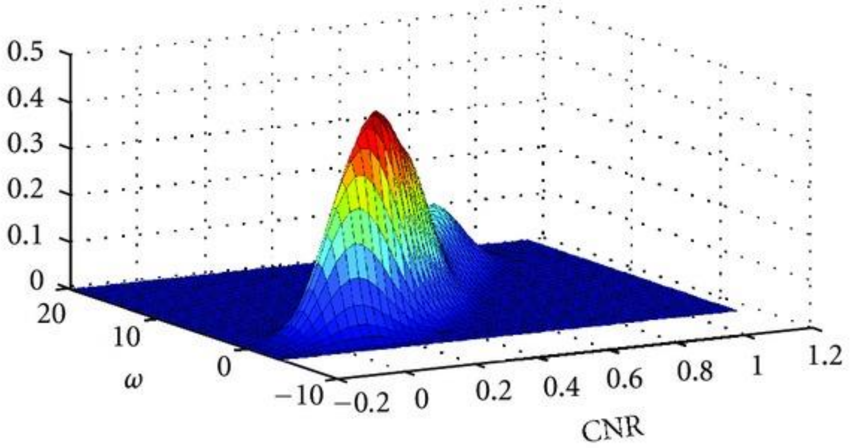




Random vectors







Data → r.v.

Relative frequency

Relative frequency is the same as experimental probability. We use relative frequency to predict probabilities from experimental data.

<u>The experiment</u>

This spinner was spun 40 times and the results recorded in this table:

Colour	Frequency
Blue	20
Yellow	10
Red	5
Green	5

<u>Relative frequency</u>

$$\boxed{\frac{\text{frequency of event}}{\text{total number of trials}}}$$

Event means one possible outcome; here, one colour on the spinner.

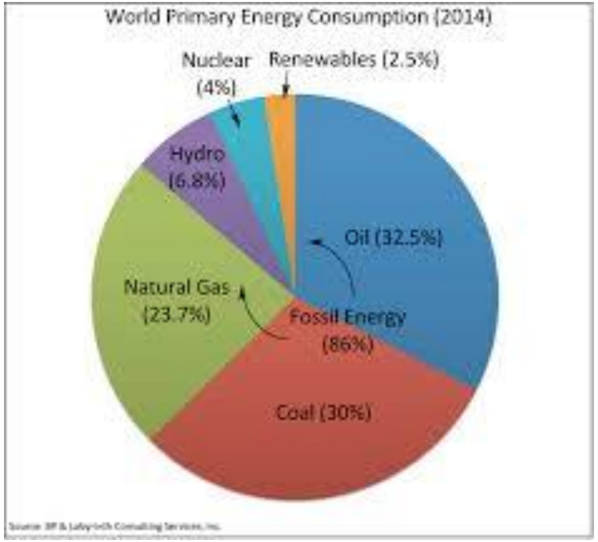
There were 20 blues recorded...

...out of 40 spins.

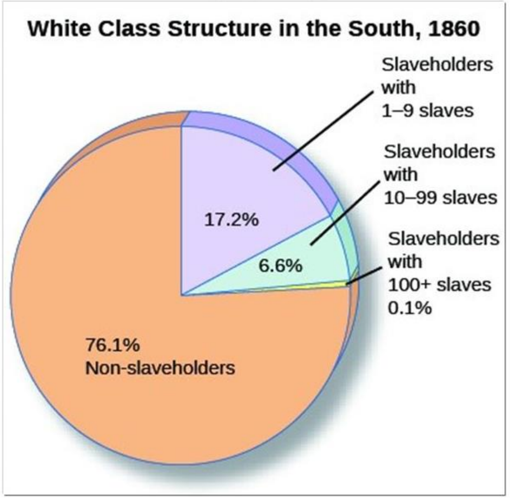
$$P(\text{blue}) = \frac{20}{40}$$

Simplify: $P(\text{blue}) = \frac{20}{40} = \frac{2}{4} = \frac{1}{2}$

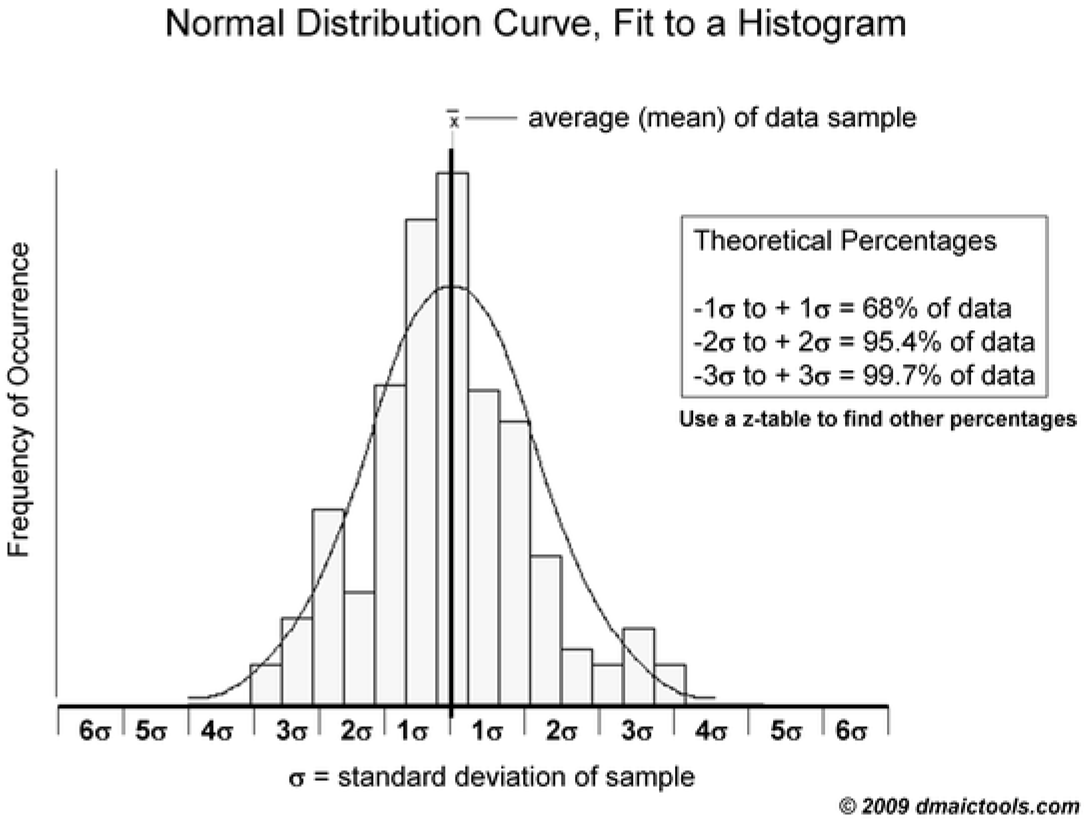
Discrete Prior distributions



Slave-Owning Population (1860)



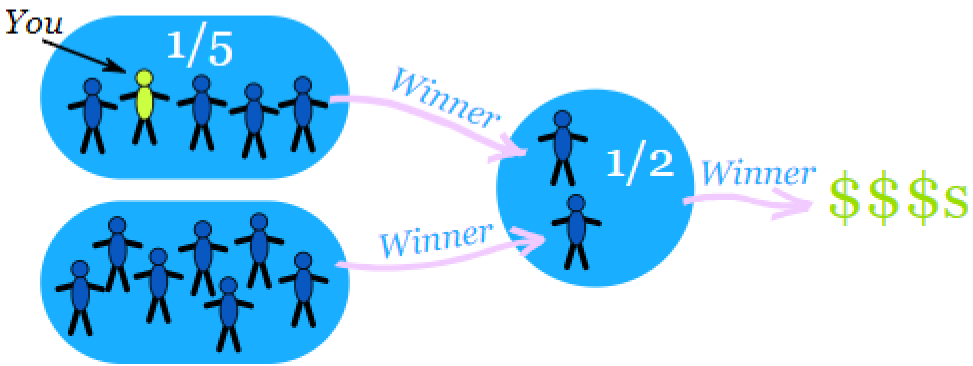
Data → r.v.


Normal Distribution Curve, Fit to a Histogram
x̄ average (mean) of data sample
Frequency of Occurrence
Theoretical Percentages
-1σ to + 1σ = 68% of data
-2σ to + 2σ = 95.4% of data
-3σ to + 3σ = 99.7% of data
Use a z-table to find other percentages
6σ 5σ 4σ 3σ 2σ 1σ 1σ 2σ 3σ 4σ 5σ 6σ
σ = standard deviation of sample
© 2009 dmaictools.com

Independent Events

Imagine there are two groups:

- A member of each group gets randomly chosen for the winners circle,
- **then** one of those gets randomly chosen to get the big money prize:



What is your chance of winnning the big prize?

Independent vs. Dependent Events

Using the bag of marbles on the left, what is the probability of pulling a black marble two times in a row? $P(black, black)$

When you put 1st marble back in (*Independent Events*)

$$\frac{2}{10} * \frac{2}{10}$$

$$\frac{1}{5} * \frac{1}{5} = \frac{1}{25}$$

When you KEEP 1st marble (*Dependent Events*)

$$\frac{2}{10} * \frac{1}{9}$$

$$\frac{1}{5} * \frac{1}{9}$$

Independent Events

The outcome of one event does not affect the outcome of the other.

If A and B are independent events then the probability of both occurring is

$$P(A \text{ and } B) = P(A) \times P(B)$$

Dependent Events

The outcome of one event affects the outcome of the other.

If A and B are dependent events then the probability of both occurring is

$$P(A \text{ and } B) = P(A) \times P(B|A)$$

Probability of B given A

Independent vs. Dependent Events

Using the bag of marbles on the left, what is the probability of pulling a black marble two times in a row? $P(black, black)$

When you put 1st marble back in (*Independent Events*)

$$\frac{2}{10} * \frac{2}{10}$$

$$\frac{1}{5} * \frac{1}{5} = \frac{1}{25}$$

When you KEEP 1st marble (*Dependent Events*)

$$\frac{2}{10} * \frac{1}{9}$$

$$\frac{1}{5} * \frac{1}{9}$$

$$P(A \text{ and } B) = P(A) \times P(B)$$

$$P(A \text{ and } B) = P(A) \times P(B|A)$$

Probability of B given A

Marginal Probabilities

$$\begin{cases} x = 1 & \text{(Rains)} \\ x = 0 & \text{(Doesn't rain)} \end{cases}$$

$Pr(x = 1) = 0.6$

$Pr(x = 0) = 0.4$

$$\begin{cases} y = 1 & \text{(Have umbrella)} \\ y = 0 & \text{(Don't have umbrella)} \end{cases}$$

$Pr(y = 1) = 0.3$

$Pr(y = 0) = 0.7$

Joint Probability

$\begin{cases} x = 1 & \text{(Rains)} \\ x = 0 & \text{(Doesn't rain)} \end{cases}$

$\begin{cases} y = 1 & \text{(Have umbrella)} \\ y = 0 & \text{(Don't have umbrella)} \end{cases}$

$Pr(x = 1) = 0.6$

$Pr(x = 0) = 0.4$

$Pr(y = 1) = 0.3$

$Pr(y = 0) = 0.7$

$$Pr(x=0) = \sum_{y=0}^{1} Pr(x=0, y)$$
$$= Pr(x=0, y=0) + Pr(x=0, y=1)$$
$$= 0.28 + 0.12 = 0.4$$

<u>Case 1 : Rains but you have an umbrella</u>

$$Pr(x = 1, y = 1) = Pr(x=1) \times Pr(y = 1)$$
$$= 0.6 \times 0.3$$
$$= 0.18$$

<u>Case 2 : Rains but you DON'T have an umbrella</u>

$$Pr(x = 1, y = 0) = Pr(x=1) \times Pr(y=0)$$
$$= 0.6 \times 0.7$$
$$= 0.42$$

Conditional Probability

Given

$x = 1$ (Rains)

$\downarrow$

What's the Probability of

$y = 1$ (Bring umbrella)

$$\begin{cases} x = 1 & \text{(Rains)} \\ x = 0 & \text{(Doesn't rain)} \end{cases}$$

$$\begin{cases} y = 1 & \text{(Have umbrella)} \\ y = 0 & \text{(Don't have umbrella)} \end{cases}$$

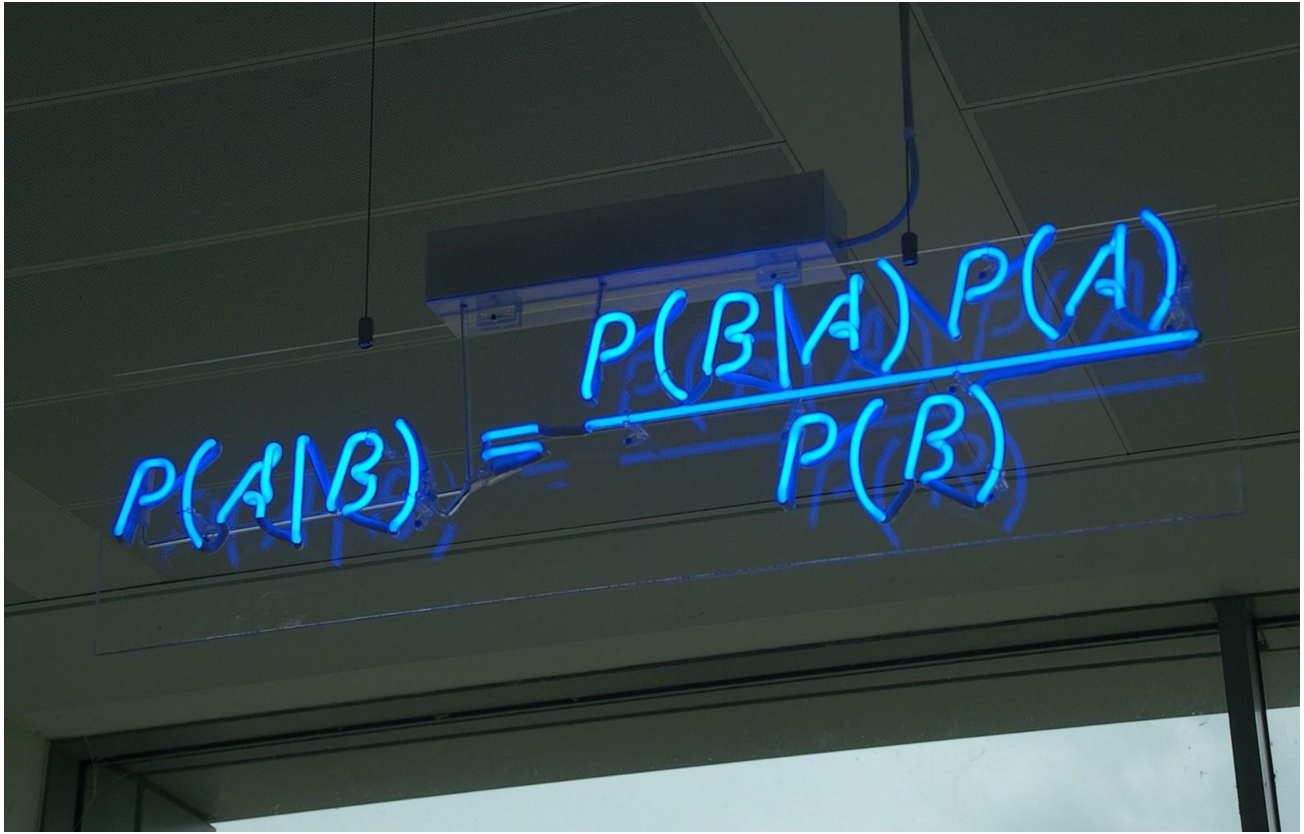
$Pr(x = 1) = 0.6$

$Pr(x = 0) = 0.4$

$Pr(y = 1) = 0.3$

$Pr(y = 0) = 0.7$

Bayes' Rule



Our first machine learning problem:

Parameter estimation:
MLE, MAP

Estimating Probabilities

Flipping a Coin

I have a coin, if I flip it, what's the probability that it will fall with the head up?

Let us flip it a few times to estimate the probability:

The estimated probability is: **3/5** "Frequency of heads"

Flipping a Coin

The estimated probability is: **3/5** "Frequency of heads"

Questions:

(1) Why frequency of heads???

(2) How good is this estimation???

(3) Why is this a machine learning problem???

We are going to answer these questions

Question (1)

Why frequency of heads???

- Frequency of heads is exactly the
  maximum likelihood estimator for this problem

- MLE has nice properties
  (interpretation, statistical guarantees, simple)

Maximum Likelihood Estimation

MLE for Bernoulli distribution

Data, D =

$$D = \{X_i\}_{i=1}^n, \; X_i \in \{\mathrm{H}, \mathrm{T}\}$$

$P(Heads) = \theta, \;\; P(Tails) = 1\text{-}\theta$

Flips are **i.i.d.**:

- **Independent** events
  - **Identically distributed** according to Bernoulli distribution

MLE: Choose θ that maximizes the probability of observed data

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

$P(x_1, x_2, \ldots x_n / \theta)$

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)$$

Independent draws

n

$$= \arg\max_{\theta} \prod_{i:X_i=H} \theta \prod_{i:X_i=T} (1-\theta)$$

Identically distributed

$$= \arg\max_{\theta} \underbrace{\theta^{\alpha_H}(1-\theta)^{\alpha_T}}_{J(\theta)}$$

n

α_H

α_T

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D|\theta)$$

Question (2)

How good is this MLE estimation???

$$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T} = \underbrace{\frac{\alpha_H}{n}}$$

$$E_n[\alpha_H] = np$$

$$E\left[\frac{\alpha_H}{n}\right] = \frac{np}{n} = p$$

$$\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right] = \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)^2\right]}_{\text{var}(\mathbf{x})} + \underbrace{\left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2}_{\text{bias}(\mathbf{x})}$$

$$\left(E_{\alpha_H}[\text{estimator}] - \underset{p}{\theta}\right)^2$$

How many flips do I need?

I flipped the coins 5 times: 3 heads, 2 tails

$$\widehat{\theta}_{MLE} = \frac{3}{5}$$

What if I flipped 30 heads and 20 tails?

$$\widehat{\theta}_{MLE} = \frac{30}{50}$$

- **Which estimator should we trust more?**
- **The more the merrier???**

Simple bound

Let θ^* be the true parameter.

For $n = \alpha_H + \alpha_T$, and $\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$

For any $\varepsilon > 0$:

Hoeffding's inequality:

$$P(|\widehat{\theta} - \theta^*| \geq \epsilon) \leq 2e^{-2n\epsilon^2}$$

Probably Approximate Correct (PAC)Learning

I want to know the coin parameter θ, within $\varepsilon = 0.1$ error with probability at least $1\text{-}\delta = 0.95$.

How many flips do I need?

$$P(|\,\widehat{\theta} - \theta^*\,| \geq \epsilon) \;\leq\; 2e^{-2n\epsilon^2} \leq \delta$$

0.05

$$e^{-2n\varepsilon^2} \leq \frac{\delta}{2}$$

$$-2n\varepsilon^2 \leq \ln\left(\frac{\delta}{2}\right)$$

$$\ln\left(\frac{2}{\delta}\right) \leq 2n\varepsilon^2 \Rightarrow n \geq \frac{1}{2\varepsilon^2}\ln\left(\frac{2}{\delta}\right)$$

Sample complexity:

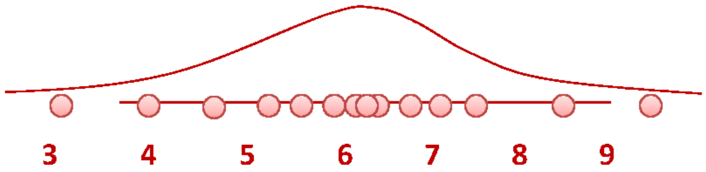
$$n \geq \frac{\ln(2/\delta)}{2\epsilon^2}$$

Question (3)

Why is this a machine learning problem???

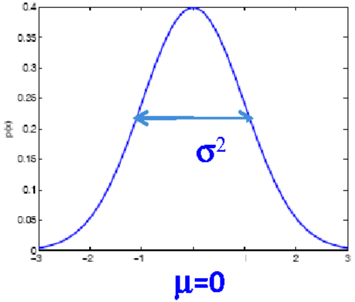
- improve their performance (accuracy of the predicted prob.)
- at some task (predicting the probability of heads)
- with experience (the more coins we flip the better we are)

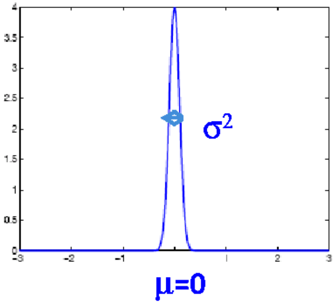
What about continuous features?



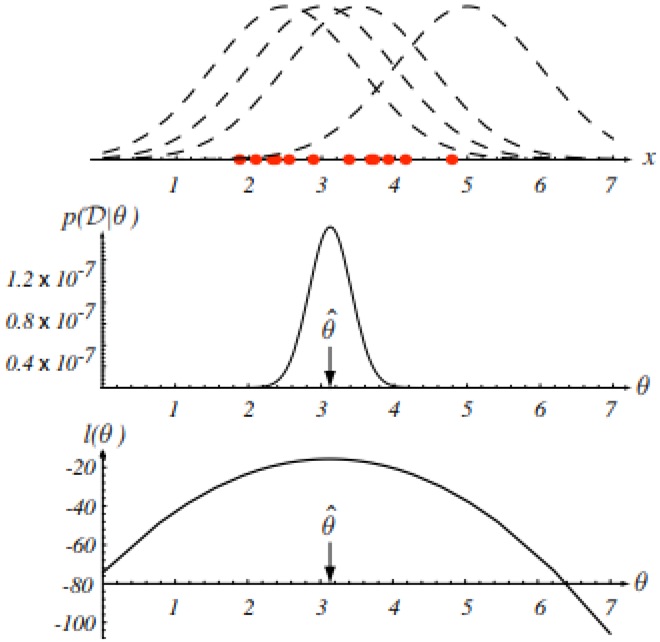
Let us try Gaussians…

$$p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = \mathcal{N}_x(\mu, \sigma)$$





Example: Maximum Likelihood Estimate of the Mean



MLE for Gaussian mean and variance

Choose $\theta = (\mu, \sigma^2)$ that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \; P(D \mid \theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)$$

Independent draws

$$= \arg\max_{\theta} \prod_{i=1}^{n} \frac{1}{2\sigma^2} e^{-(X_i-\mu)^2/2\sigma^2}$$

Identically distributed

$$= \arg \max_{\theta=(\mu,\sigma^2)} \underbrace{\frac{1}{2\sigma^2} e^{-\sum_{i=1}^{n}(X_i-\mu)^2/2\sigma^2}}_{J(\theta)}$$

MLE for Gaussian mean and variance

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

$$\widehat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \widehat{\mu})^2$$

Note: MLE for the variance of a Gaussian is **biased**

[Expected result of estimation is **not** the true parameter!]

Unbiased variance estimator: $\widehat{\sigma}^2_{unbiased} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \widehat{\mu})^2$

$\mathbb{E}[\hat{\sigma}^2_{MLE}] \neq \sigma^2 \quad \mathbb{E}[\hat{\sigma}^2_{UB}] = \sigma^2$

What about prior knowledge?

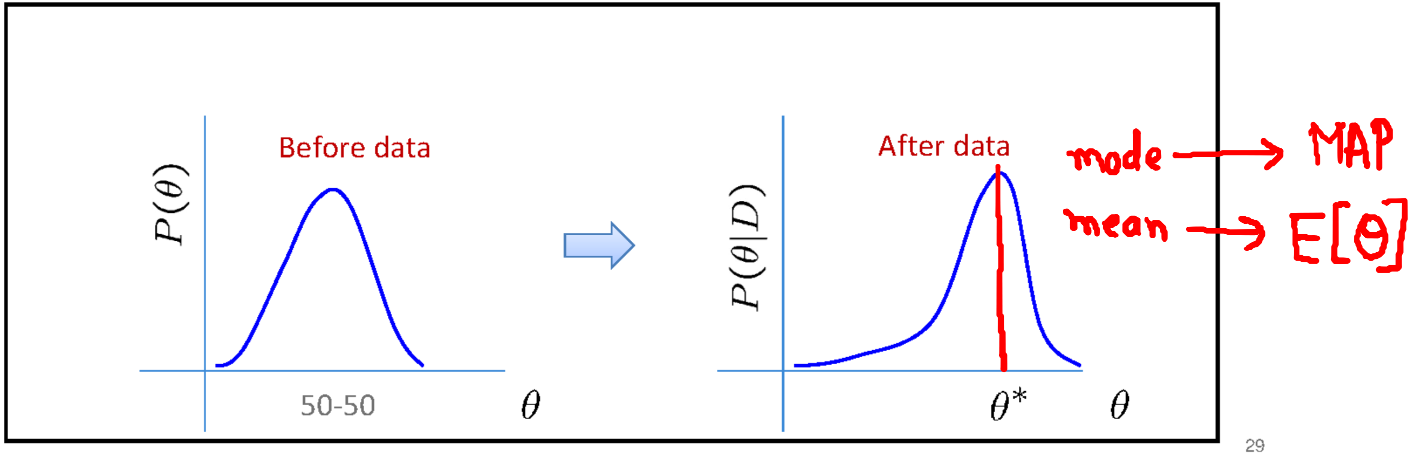
(MAP Estimation)

What about prior knowledge?

We know the coin is "close" to 50-50. What can we do now?

The Bayesian way...

Rather than estimating a single θ, we obtain a distribution over possible values of θ



Prior distribution

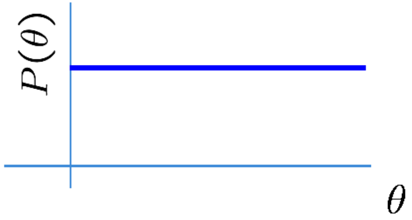
$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

What prior? What distribution do we want for a prior?

- Represents expert knowledge (philosophical approach)
- Simple posterior form (engineer's approach)

Uninformative priors:

- Uniform distribution



Conjugate priors:

- Closed-form representation of posterior
- $P(\theta)$ and $P(\theta|D)$ have the same form

In order to proceed we will need:

Bayes Rule

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

Chain Rule & Bayes Rule

Chain rule:

$$P(X,Y) = P(X|Y)P(Y) = P(Y|X)P(X)$$

Bayes rule:

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

Bayes rule is important for reverse conditioning.

Bayesian Learning

- Use Bayes rule:

$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})}$$

- Or equivalently:

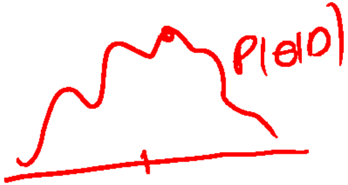
$$\underbrace{P(\theta \mid \mathcal{D})}_{\text{posterior}} \propto \underbrace{P(\mathcal{D} \mid \theta)}_{\text{likelihood}}\underbrace{P(\theta)}_{\text{prior}}$$

MLE vs. MAP

- Maximum Likelihood estimation (MLE)

  Choose value that maximizes the probability of observed data $\widehat{\theta}_{MLE} = \arg\max_{\theta} P(D|\theta)$

$\mathbb{E}[\theta|D]$



- Maximum *a posteriori* (MAP) estimation

  Choose value that is most probable given observed data and prior belief

$$\widehat{\theta}_{MAP} = \arg\max_{\theta} P(\theta|D)$$
$$= \arg\max_{\theta} P(D|\theta)P(\theta)$$

When is MAP same as MLE?

MAP estimation for Binomial distribution

Coin flip problem: Likelihood is Binomial

$$P(\mathcal{D} \mid \theta) = \binom{n}{\alpha_H} \theta^{\alpha_H} (1-\theta)^{\alpha_T}$$

If the prior is Beta distribution,

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$

$\Rightarrow$ posterior is Beta distribution

Beta function: $B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1}\, dt$

MAP estimation for Binomial distribution

Likelihood is Binomial: $P(\mathcal{D} \mid \theta) = \binom{n}{\alpha_H} \theta^{\alpha_H}(1-\theta)^{\alpha_T}$

Prior is Beta distribution: $P(\theta) = \frac{\theta^{\beta_H-1}(1-\theta)^{\beta_T-1}}{B(\beta_H,\beta_T)} \sim Beta(\beta_H,\beta_T)$

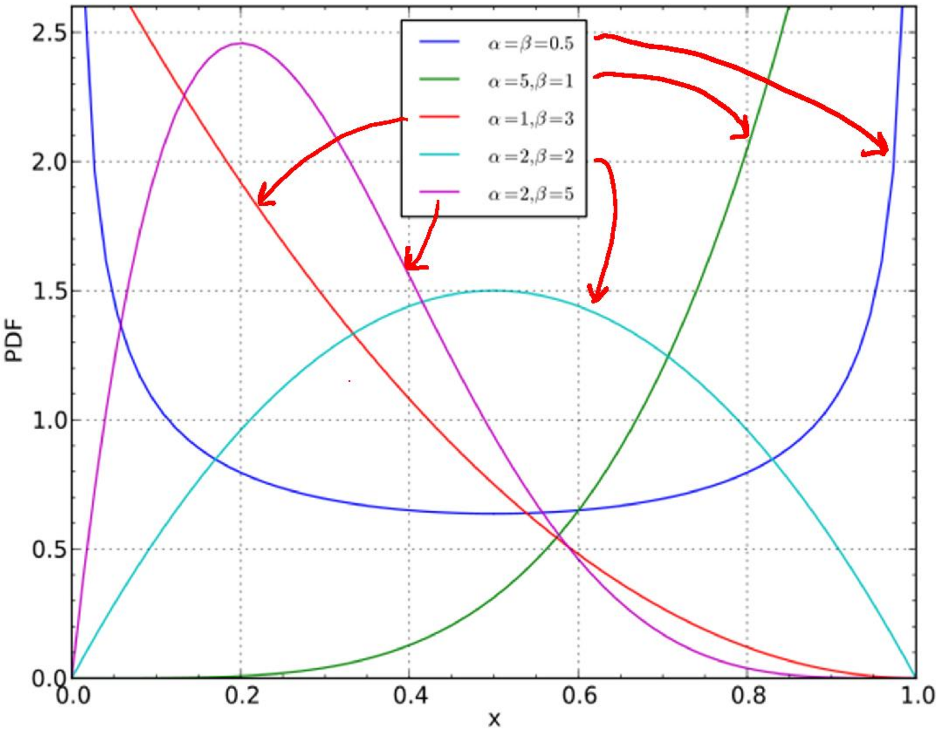
$\Rightarrow$ posterior is Beta distribution

$$P(\theta|D) \sim Beta(\beta_H+\alpha_H, \beta_T+\alpha_T)$$

P(θ) and P(θ|D) have the same form! [Conjugate prior]

$$\widehat{\theta}_{MAP} = \arg\max_{\theta} \; P(\theta \mid D) = \arg\max_{\theta} \; P(D \mid \theta)P(\theta)$$
$$= \frac{\alpha_H+\beta_H-1}{\alpha_H+\beta_H+\alpha_T+\beta_T-2}$$

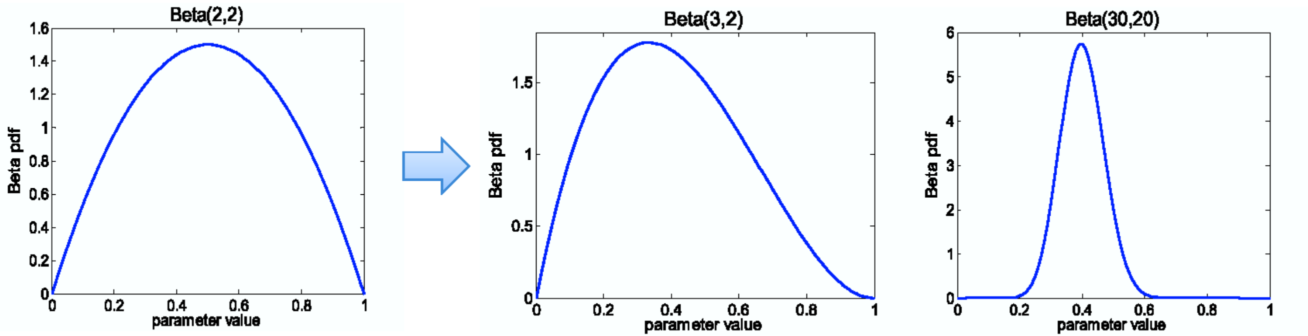
Beta distribution



More concentrated as values of α, β increase

Beta conjugate prior

$P(\theta) \sim Beta(\beta_H, \beta_T)$ $\qquad$ $P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$



As $n = \alpha_H + \alpha_T$ increases

As we get more samples, effect of prior is “washed out”

From Binomial to Multinomial

Example: Dice roll problem (6 outcomes instead of 2)

Likelihood is ~ Multinomial($\theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$)

$$P(\mathcal{D} \mid \theta) = \theta_1^{\alpha_1} \theta_2^{\alpha_2} \ldots \theta_k^{\alpha_k}$$

If prior is Dirichlet distribution,

$$P(\theta) = \frac{\prod_{i=1}^{k} \theta_i^{\beta_i - 1}}{B(\beta_1, \ldots, \beta_k)} \sim \text{Dirichlet}(\beta_1, \ldots, \beta_k)$$

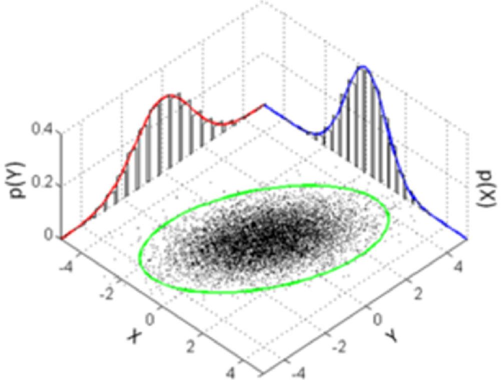
Then posterior is Dirichlet distribution

$$P(\theta|D) \sim \text{Dirichlet}(\beta_1 + \alpha_1, \ldots, \beta_k + \alpha_k)$$

For Multinomial, conjugate prior is Dirichlet distribution.

http://en.wikipedia.org/wiki/Dirichlet_distribution

Conjugate prior for Gaussian?



$$(2\pi)^{-\frac{k}{2}}|\boldsymbol\Sigma|^{-\frac{1}{2}}\, e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol\mu)'\boldsymbol\Sigma^{-1}(\mathbf{x}-\boldsymbol\mu)},$$

Conjugate prior on mean: **Gaussian**

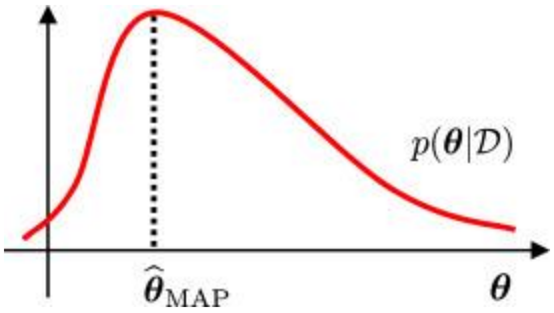
Conjugate prior on covariance matrix: **Inverse Wishart**

$$\frac{\left|\boldsymbol\Psi\right|^{\frac{\nu}{2}}}{2^{\frac{\nu p}{2}}\Gamma_p(\frac{\nu}{2})}\left|\mathbf{X}\right|^{-\frac{\nu+p+1}{2}} e^{-\frac{1}{2}\operatorname{tr}(\boldsymbol\Psi\mathbf{X}^{-1})}$$

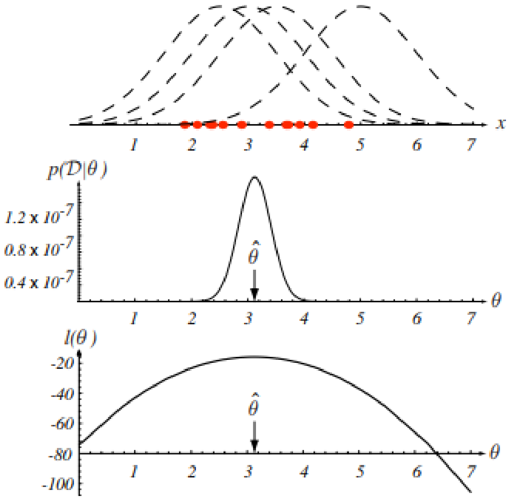
Bayesians vs.Frequentists

$$
\begin{aligned}
\widehat{\Theta}_{MAP} &= \underset{\Theta}{\operatorname{argmax}} \ prob(\Theta|\mathcal{X}) \\
&= \underset{\Theta}{\operatorname{argmax}} \ \frac{prob(\mathcal{X}|\Theta) \ \cdot \ prob(\Theta)}{prob(\mathcal{X})} \\
&= \underset{\Theta}{\operatorname{argmax}} \ prob(\mathcal{X}|\Theta) \ \cdot \ prob(\Theta) \\
&= \underset{\Theta}{\operatorname{argmax}} \ \prod_{\mathbf{x}_i \in \mathcal{X}} prob(\mathbf{x}_i|\Theta) \ \cdot \ prob(\Theta)
\end{aligned}
$$

$$
\begin{aligned}
\widehat{\Theta}_{MAP} &= \underset{\Theta}{\operatorname{argmax}}\ prob(\Theta|\mathcal{X}) \\
&= \underset{\Theta}{\operatorname{argmax}}\ \frac{prob(\mathcal{X}|\Theta)\ \cdot\ prob(\Theta)}{prob(\mathcal{X})} \\
&= \underset{\Theta}{\operatorname{argmax}}\ prob(\mathcal{X}|\Theta)\ \cdot\ prob(\Theta) \\
&= \underset{\Theta}{\operatorname{argmax}}\ \prod_{\mathbf{x}_i\in\mathcal{X}} prob(\mathbf{x}_i|\Theta)\ \cdot\ prob(\Theta)
\end{aligned}
$$

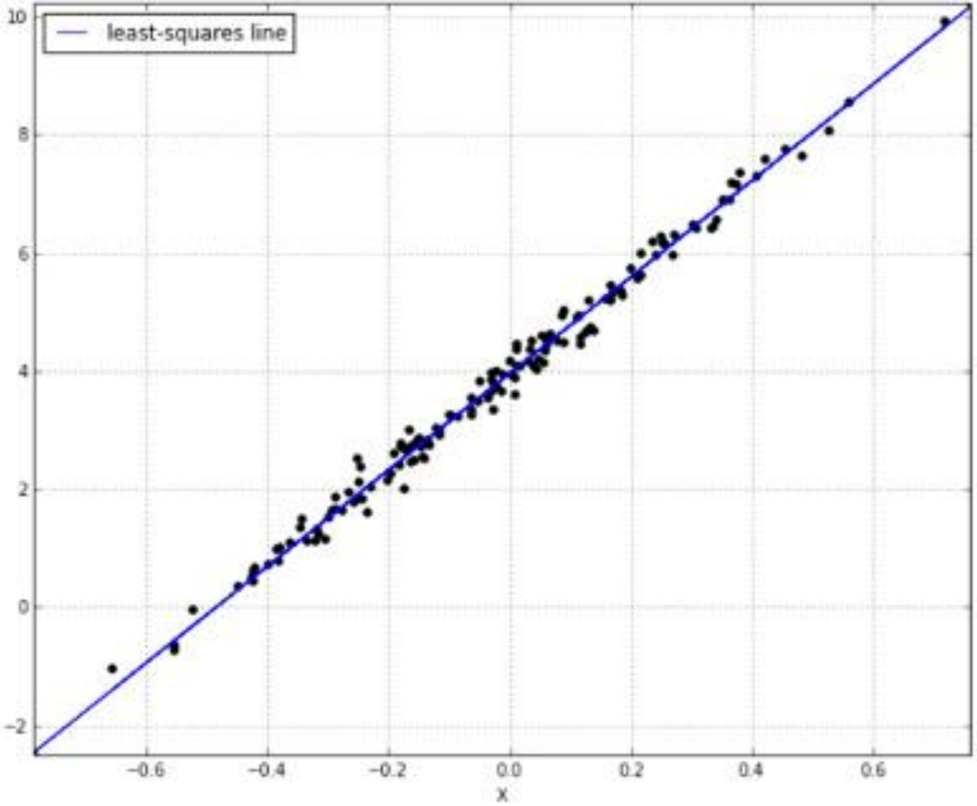


$$
\widehat{\Theta}_{MAP}=\underset{\Theta}{\operatorname{argmax}}\left(\sum_{\mathbf{x}_i\in\mathcal{X}} log\ prob(\mathbf{x}_i|\Theta)\ +\ log\ prob(\Theta)\right)
$$



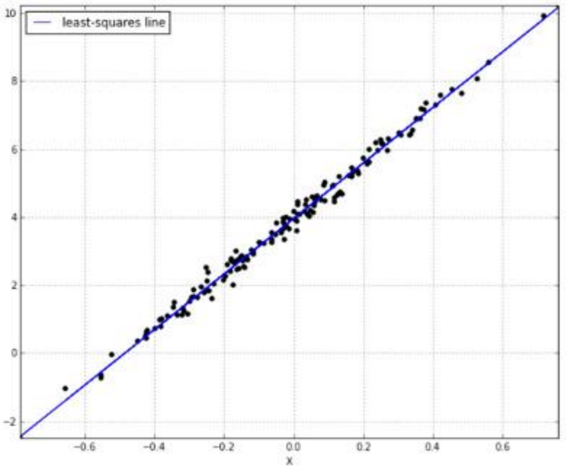
$$y_i = \underbrace{\theta^T x_i}_{\hat{y}_i} + \epsilon_i$$

$$\arg\min_\theta \sum_{i=1}^n (y_i - \hat{y}_i)^2$$



y = 4 + 8x + noise

Linear Regression Param Estimation - A Probabilistic Perspective



$$y_i = \theta^T x_i + \epsilon_i \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$$X = \mu + Y$$
$$Y \sim \mathcal{N}(0, \sigma^2)$$

$$\Rightarrow \quad X \sim \mathcal{N}(\mu, \sigma^2) \qquad \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$y_i \sim \mathcal{N}(\theta^T x_i, \sigma^2)$$

$p(Y_1 = y_1 | x, \theta)$

$$\text{likelihood} = p(y_1 | x_1, \theta) * p(y_2 | x_2, \theta) \ldots * p(y_n | x_n, \theta)$$

$(X_1, c_1)(X_1, c_2) \cdots (X_n, c_n)$

$$\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i - \theta^T x_i)^2}{2\sigma^2}}$$

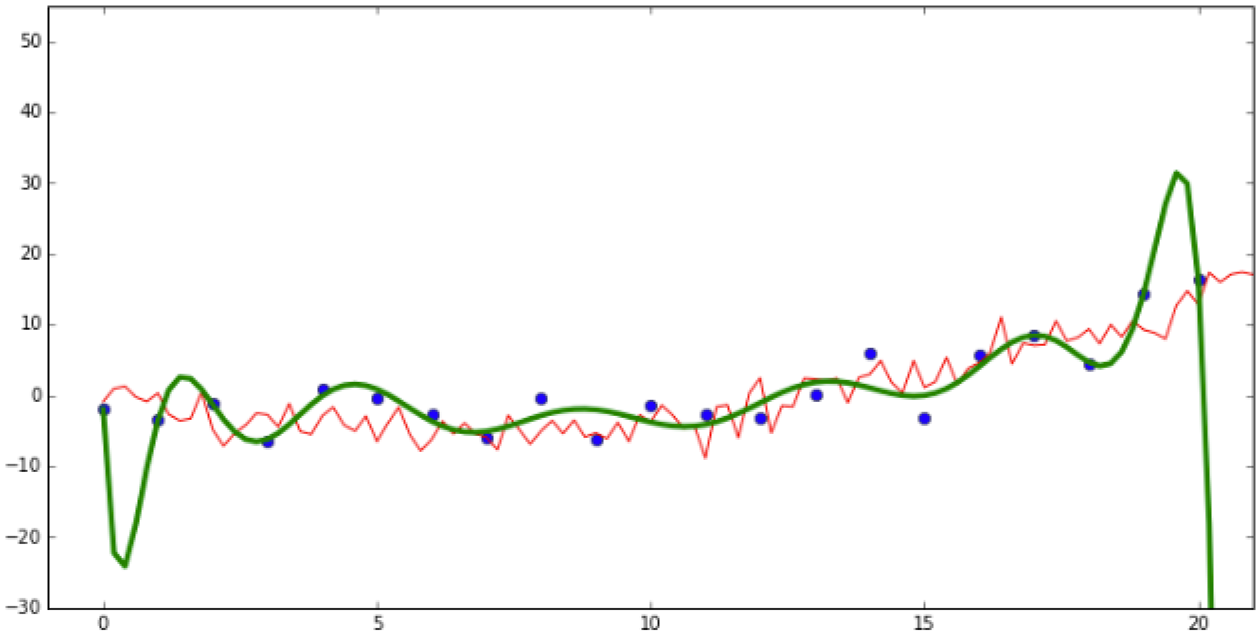
$p(c = a | x)$

$$\sigma^2 2\pi^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2} \sum_1^n (y_i - \theta^T x_i)^2}$$

$$\arg\min_\theta \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

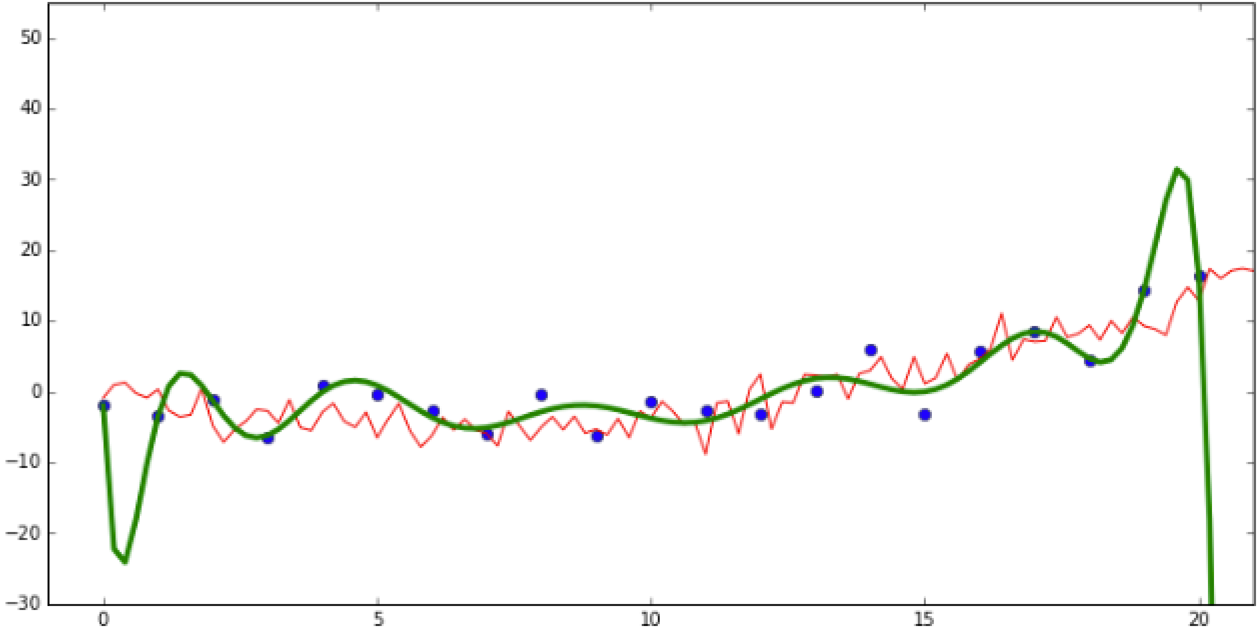
Minimizing the residual sum of squares is equivalent to maximizing the likelihood of the data

$y = ax + bx^2$ Actual



14 degree poly
(hypothesis set)

$y = \underline{a}x + \underline{b}x^2$ Actual



14 degree poly
(hypothesis set)

$y = ax + bx^2 + \underbrace{\epsilon}$

$\epsilon \sim N(0, \tau)$

$X \sim N_1(0, \sigma^2)$

$Y = X + c$

$Y \sim N(c, \sigma^2)$

$P(y_1 | \mu, \sigma^2) \cdot P(y_2 | \mu, \sigma^2) \cdots$

$$\arg\min_{\theta} \sum_1^n (y_i - \hat{y}_i)^2 + \lambda \sum_1^p \theta^2$$

Ridge (L2) Regression

$$\text{posterior} = \frac{\text{likelihood x prior}}{\text{marginal likelihood}}$$

$$\arg\min_{\theta} \sum_1^n (y_i - \hat{y}_i)^2 + \lambda \sum_1^p \theta^2$$

$$\text{posterior} \propto \text{likelihood x prior}$$

$\theta \sim N(0, \tau^2)$

$$\sigma^2 2\pi^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_1^n (y_i - \hat{y}_i)^2} \times \tau^2 2\pi^{-\frac{p}{2}} e^{-\frac{1}{2\tau^2}\sum_1^p \theta^2}$$

$$e^{-\frac{1}{2\sigma^2}\sum_1^n (y_i - \hat{y}_i)^2 - \frac{1}{2\tau^2}\sum_1^p \theta^2} \times \sigma^2 2\pi^{-\frac{n}{2}} \times \tau^2 2\pi^{-\frac{p}{2}}$$

$$\arg\max_{\theta} \; e^{-\frac{1}{2\sigma^2}\sum_1^n (y_i - \hat{y}_i)^2 - \frac{1}{2\tau^2}\sum_1^p \theta^2}$$

$$\arg\max_{\theta} - \frac{1}{2\sigma^2}\sum_1^n (y_i - \hat{y}_i)^2 - \frac{1}{2\tau^2}\sum_1^p \theta^2$$

$$\arg\max_{\theta} - 1\left(\sum_1^n (y_i - \hat{y}_i)^2 + \frac{\sigma^2}{\tau^2}\sum_1^p \theta^2\right)$$

$$\arg\min_{\theta} \sum_1^n (y_i - \hat{y}_i)^2 + \frac{\sigma^2}{\tau^2}\sum_1^p \theta^2$$

$$\text{posterior} = \frac{\text{likelihood x prior}}{\text{marginal likelihood}}$$

$$\arg\min_{\theta} \sum_1^n (y_i - \hat{y}_i)^2 + \lambda \sum_1^p \theta^2$$

$$\text{posterior} \propto \text{likelihood x prior}$$

$$\sigma^2 2\pi^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_1^n (y_i - \hat{y}_i)^2} \times \tau^2 2\pi^{-\frac{p}{2}} e^{-\frac{1}{2\tau^2}\sum_1^p \theta^2}$$

$$e^{-\frac{1}{2\sigma^2}\sum_1^n (y_i - \hat{y}_i)^2 - \frac{1}{2\tau^2}\sum_1^p \theta^2} \times \sigma^2 2\pi^{-\frac{n}{2}} \times \tau^2 2\pi^{-\frac{p}{2}}$$

$$\arg\max_{\theta} \; e^{-\frac{1}{2\sigma^2}\sum_1^n (y_i - \hat{y}_i)^2 - \frac{1}{2\tau^2}\sum_1^p \theta^2}$$

$$\arg\max_{\theta} -\frac{1}{2\sigma^2}\sum_1^n (y_i - \hat{y}_i)^2 - \frac{1}{2\tau^2}\sum_1^p \theta^2$$

$$\arg\max_{\theta} -1\left(\sum_1^n (y_i - \hat{y}_i)^2 + \frac{\sigma^2}{\tau^2}\sum_1^p \theta^2\right)$$

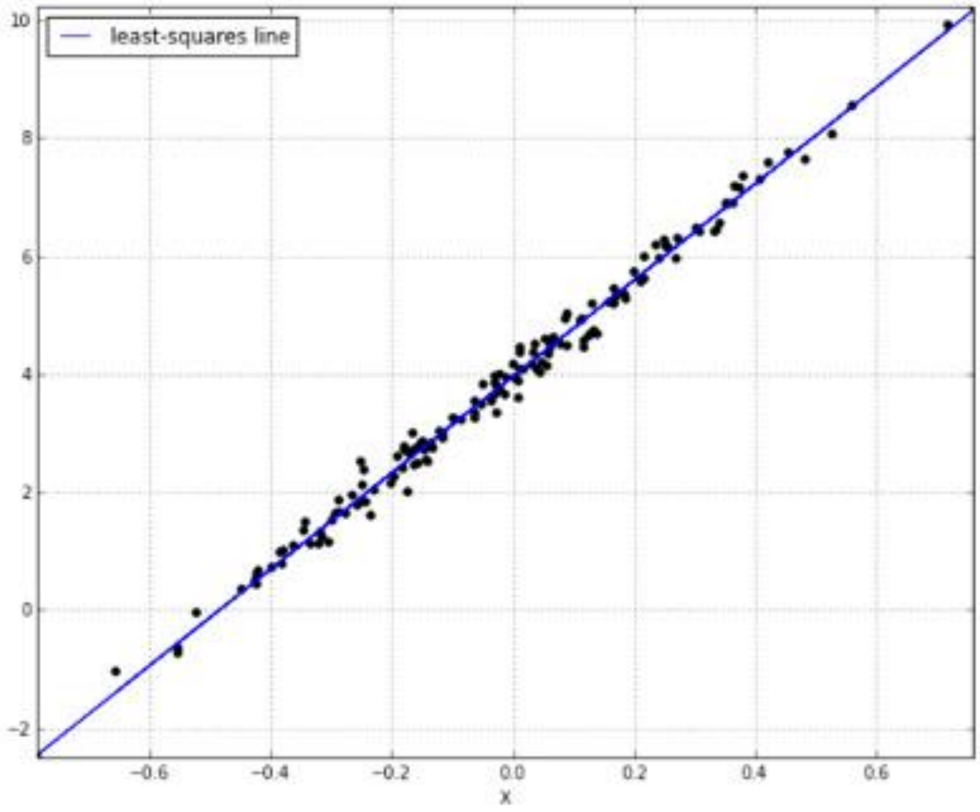
$$\arg\min_{\theta} \sum_1^n (y_i - \hat{y}_i)^2 + \frac{\sigma^2}{\tau^2}\sum_1^p \theta^2$$

Ridge regression = MAP estimate with a zero-mean Gaussian prior, with λ proportional to τ^2.

- Lower variance on the prior ➔ Higher λ value in the ridge regression solution.

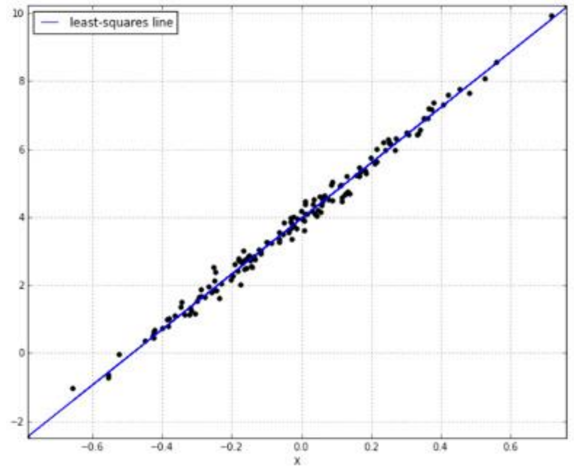
$$y_i = \underbrace{\theta^T x_i}_{\hat{y}_i} + \epsilon_i$$

$$\arg\min_{\theta} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$



y = 4 + 8x + noise

Linear Regression Param Estimation - A Probabilistic Perspective



$$y_i = \theta^T x_i + \epsilon_i \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$$X = \mu + Y$$
$$Y \sim \mathcal{N}(0, \sigma^2)$$

$$X \sim \mathcal{N}(\mu, \sigma^2) \qquad \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$y_i \sim \mathcal{N}(\theta^T x_i, \sigma^2)$$

$p(Y_1 = y_1 | x, \theta)$

$$\text{likelihood} = p(\underline{y_1}|x_1, \theta) * p(y_2|x_2, \theta) \ldots * p(y_n|x_n, \theta)$$

$(X_1, c_1)(X_1, c_2) \cdots (X_n, c_n)$

$$\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i - \theta^T x_i)^2}{2\sigma^2}}$$

$p(c = a | x)$

$$\sigma^2 2\pi^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2} \sum_1^n (y_i - \theta^T x_i)^2}$$

$$\arg\min_{\theta} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Minimizing the residual sum of squares is equivalent to maximizing the likelihood of the data

$x_i, y_i, \sigma^2, n, \theta$

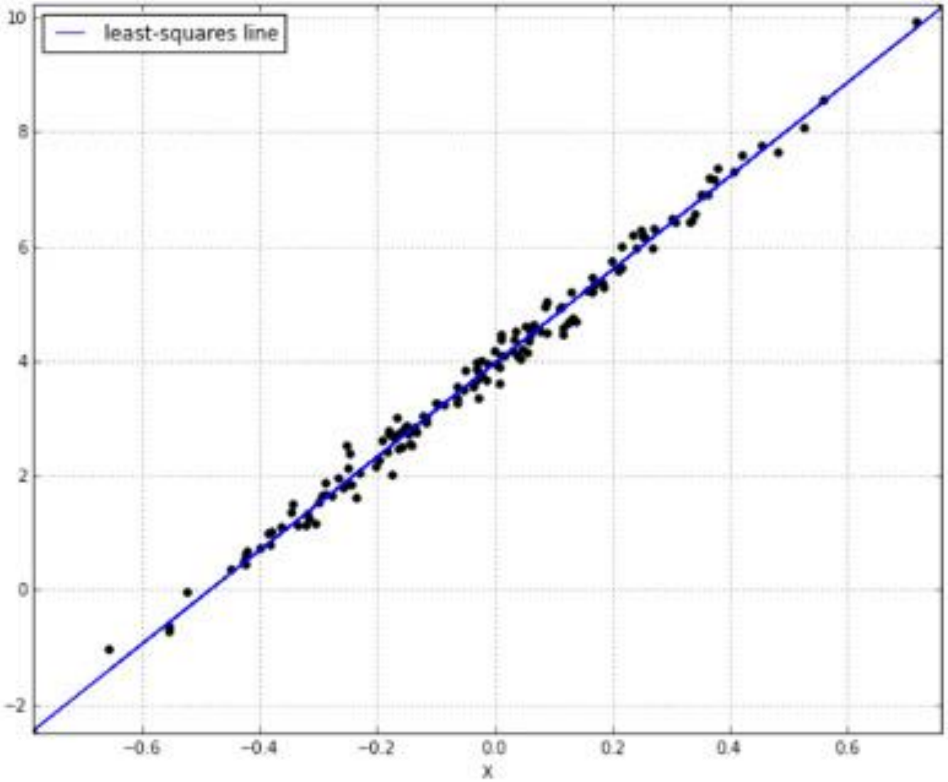
$-3, 3, 0.01$

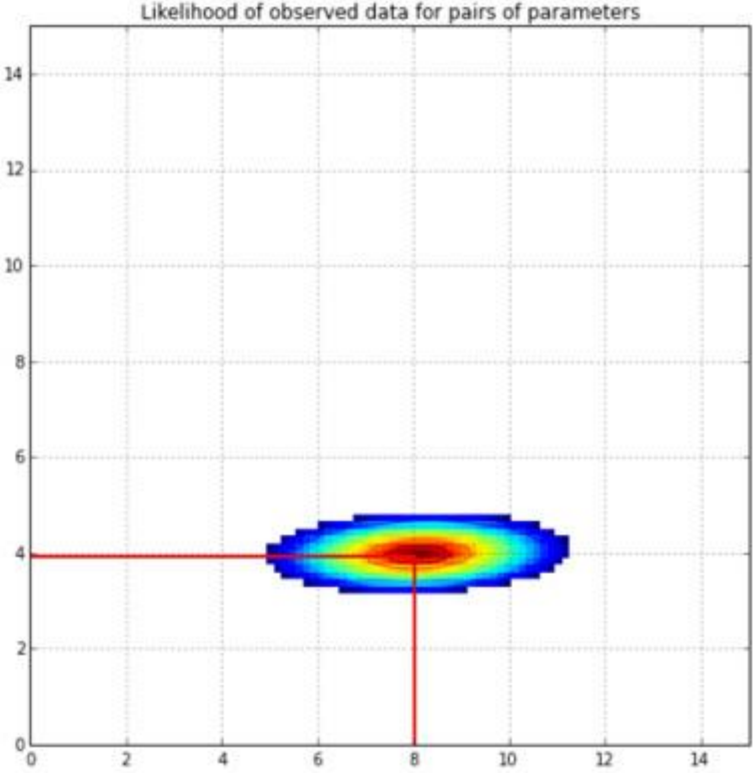
$\theta_1:$

$$L(\theta) = \sigma^2 2\pi^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_1^n (y_i - \theta^T x_i)^2}$$

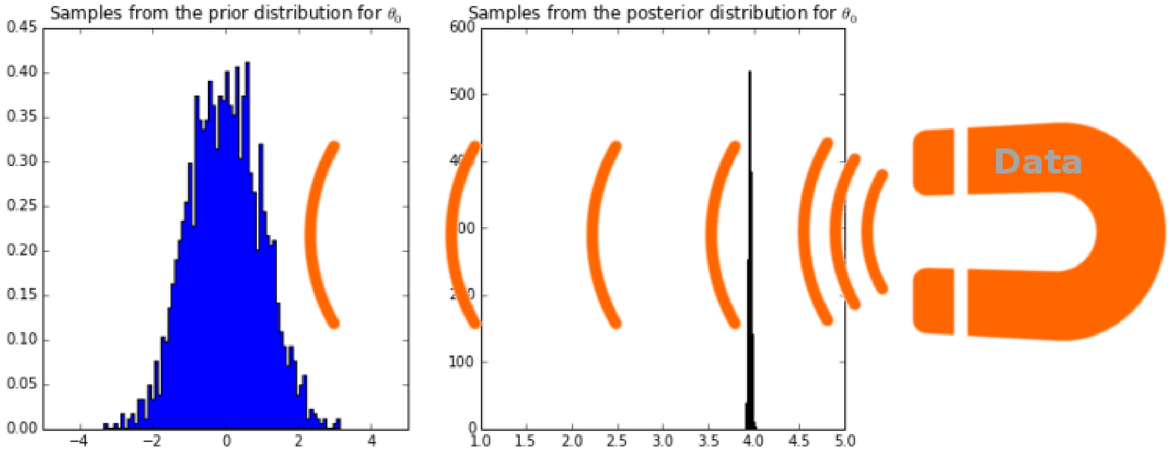
$\theta_2:$

y = 4 + 8x + noise





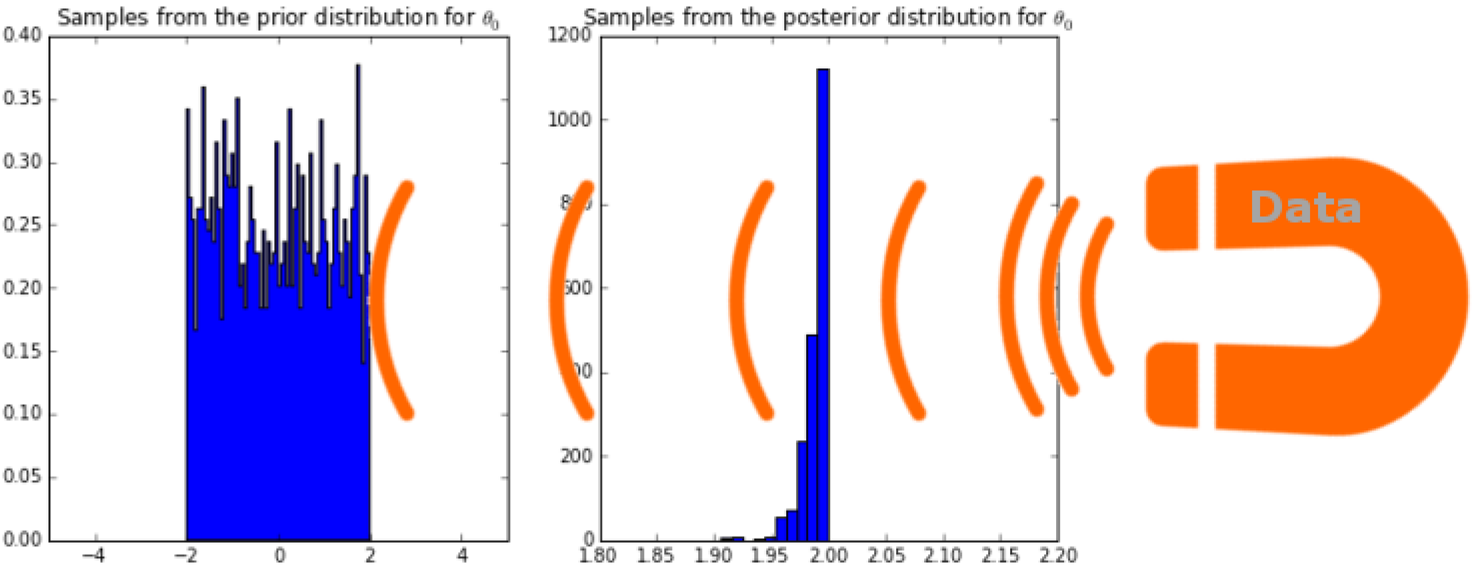
Going Bayesian – Introduce parameter priors



data is like a magnet that attracts probability mass

$$
\begin{aligned}
\widehat{\Theta}_{MAP} &= \underset{\Theta}{\text{argmax}}\ prob(\Theta|\mathcal{X}) \\
&= \underset{\Theta}{\text{argmax}}\ \frac{prob(\mathcal{X}|\Theta)\ \cdot\ prob(\Theta)}{prob(\mathcal{X})} \\
&= \underset{\Theta}{\text{argmax}}\ prob(\mathcal{X}|\Theta)\ \cdot\ prob(\Theta) \\
&= \underset{\Theta}{\text{argmax}}\ \prod_{\mathbf{x}_i \in \mathcal{X}} prob(\mathbf{x}_i|\Theta)\ \cdot\ prob(\Theta)
\end{aligned}
$$

Choosing a sensible prior is important !

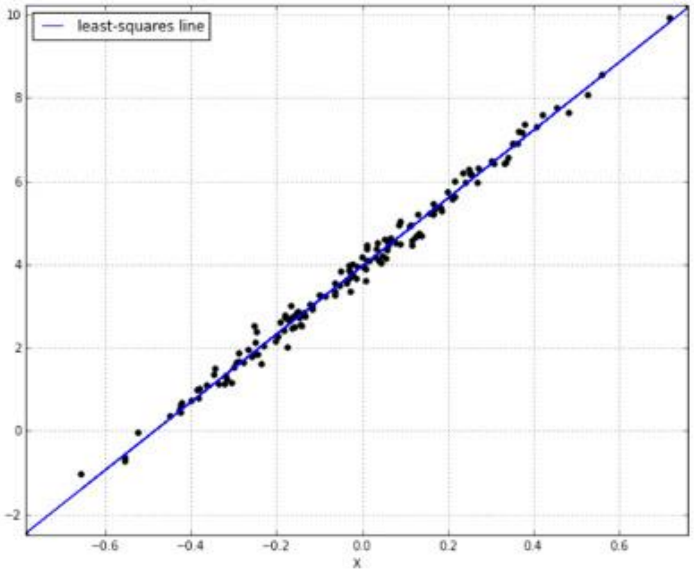


$$
\begin{aligned}
\widehat{\Theta}_{MAP} &= \underset{\Theta}{\text{argmax}}\ prob(\Theta|\mathcal{X}) \\
&= \underset{\Theta}{\text{argmax}}\ \frac{prob(\mathcal{X}|\Theta)\ \cdot\ prob(\Theta)}{prob(\mathcal{X})} \\
&= \underset{\Theta}{\text{argmax}}\ prob(\mathcal{X}|\Theta)\ \cdot\ prob(\Theta) \\
&= \underset{\Theta}{\text{argmax}}\ \prod_{\mathbf{x}_i \in \mathcal{X}} prob(\mathbf{x}_i|\Theta)\ \cdot\ prob(\Theta)
\end{aligned}
$$

$$\hat{y}^* = [\text{MAP estimate of } \theta_0] + [\text{MAP estimate of } \theta_1] x^*$$

Plug in MAP estimates ?

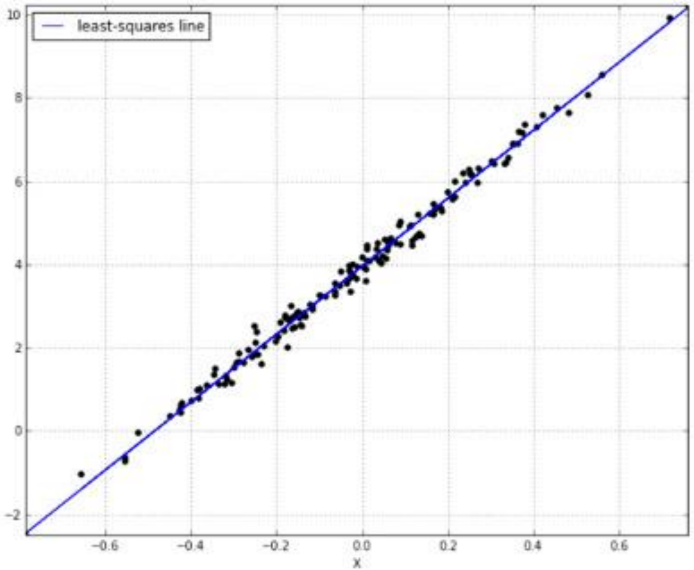
$$\begin{aligned}
\widehat{\Theta}_{MAP} &= \underset{\Theta}{\text{argmax}}\ prob(\Theta|\mathcal{X}) \\
&= \underset{\Theta}{\text{argmax}}\ \frac{prob(\mathcal{X}|\Theta)\ \cdot\ prob(\Theta)}{prob(\mathcal{X})} \\
&= \underset{\Theta}{\text{argmax}}\ prob(\mathcal{X}|\Theta)\ \cdot\ prob(\Theta) \\
&= \underset{\Theta}{\text{argmax}}\ \prod_{\mathbf{x}_i \in \mathcal{X}} prob(\mathbf{x}_i|\Theta)\ \cdot\ prob(\Theta)
\end{aligned}$$



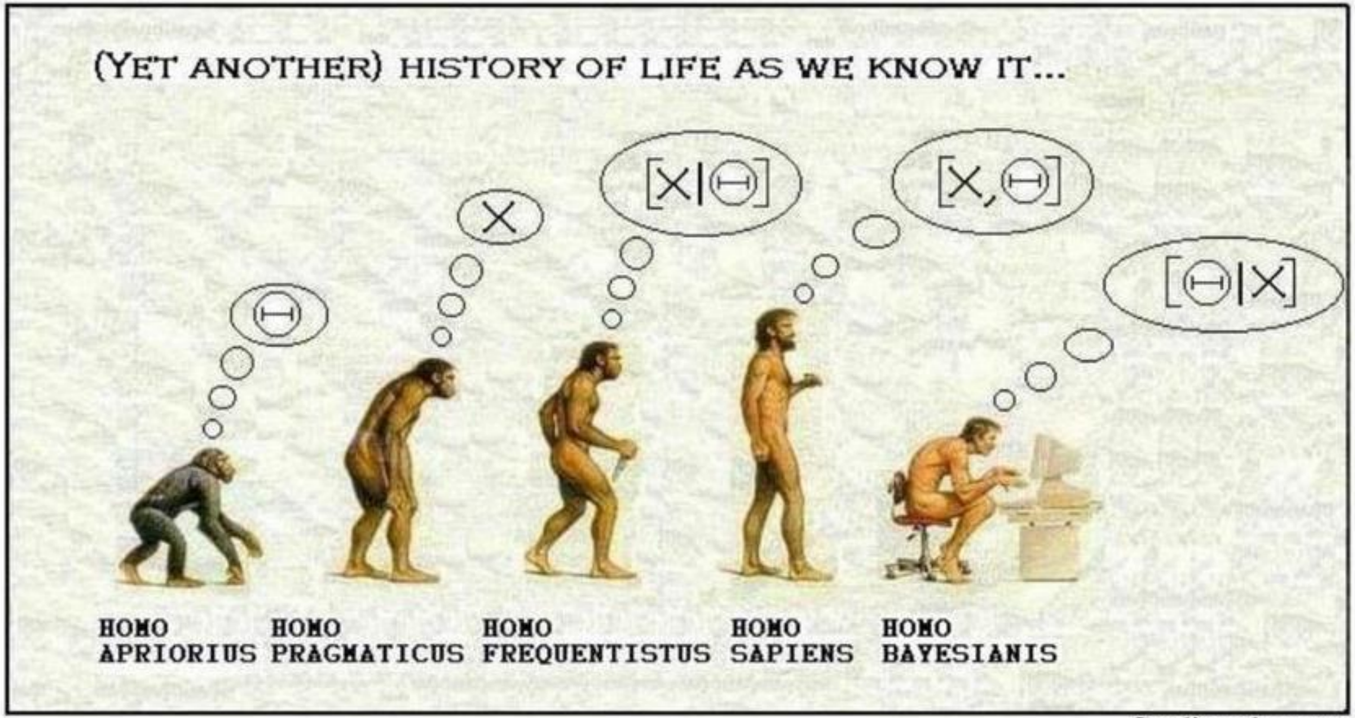
$$\hat{y^*} = [\text{MAP estimate of } \theta_0] + [\text{MAP estimate of } \theta_1] x^*$$

Plug in MAP estimates ?

$$
\begin{aligned}
\widehat{\Theta}_{MAP} &= \underset{\Theta}{\operatorname{argmax}}\ prob(\Theta|\mathcal{X}) \\
&= \underset{\Theta}{\operatorname{argmax}}\ \frac{prob(\mathcal{X}|\Theta)\ \cdot\ prob(\Theta)}{prob(\mathcal{X})} \\
&= \underset{\Theta}{\operatorname{argmax}}\ prob(\mathcal{X}|\Theta)\ \cdot\ prob(\Theta) \\
&= \underset{\Theta}{\operatorname{argmax}}\ \prod_{\mathbf{x}_i \in \mathcal{X}} prob(\mathbf{x}_i|\Theta)\ \cdot\ prob(\Theta)
\end{aligned}
$$



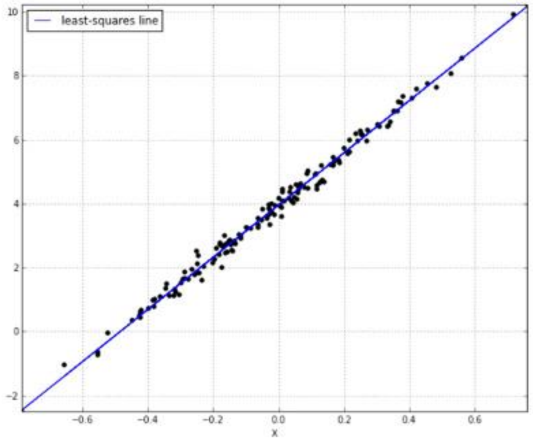
$$\hat{y^*} = [\text{MAP estimate of } \theta_0] + [\text{MAP estimate of } \theta_1]x^*$$

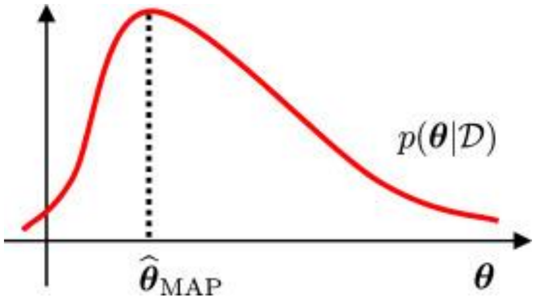


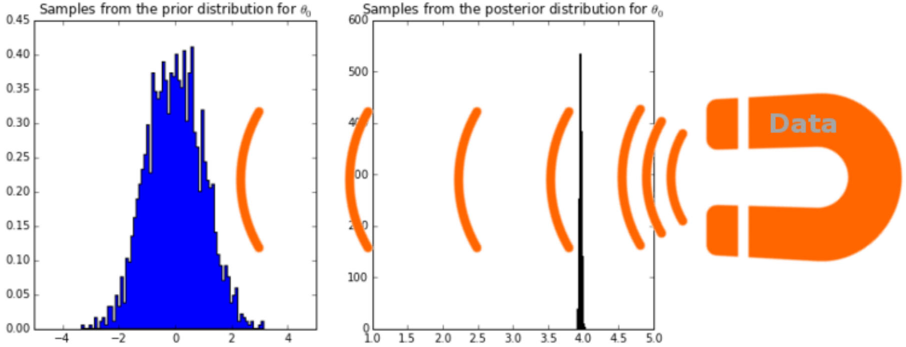
Credit:unknown

Go Full Bayesian !

$$y_i = \theta^T x_i + \epsilon_i$$





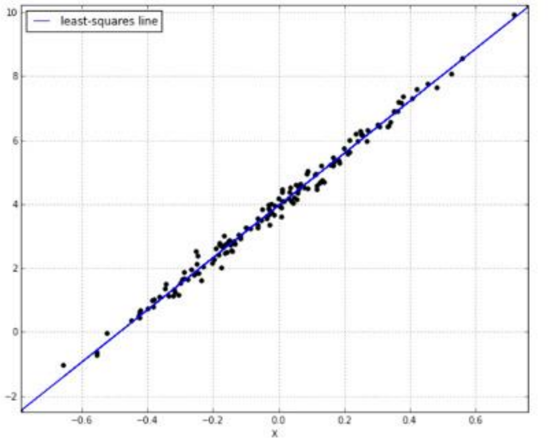


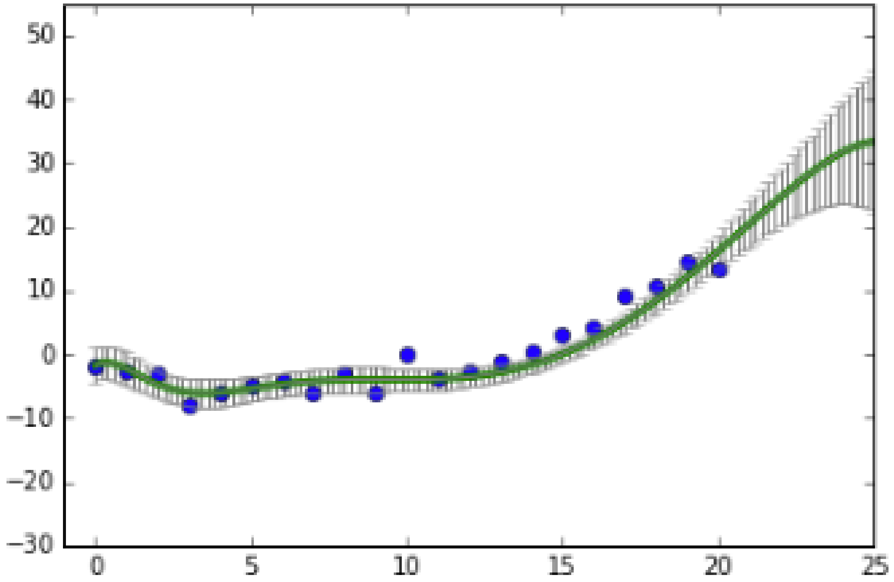
training data

$$p(y^*|x^*,X,y) = \int_\theta p(y^*|x^*,\theta)p(\theta|X,y)d\theta$$

We get a *probability distribution* for the outcome y∗ at each x∗

Go Full Bayesian !

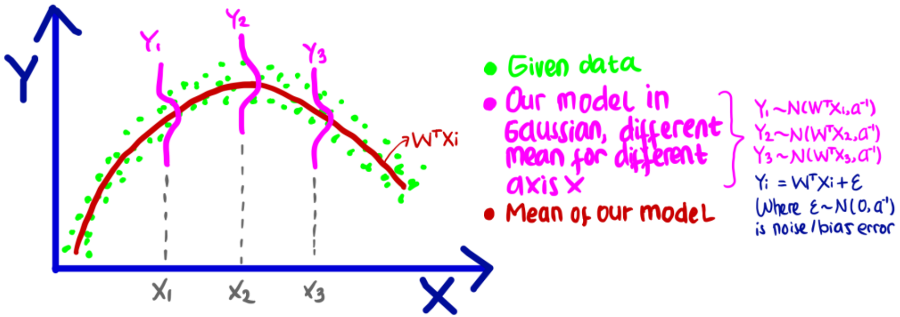


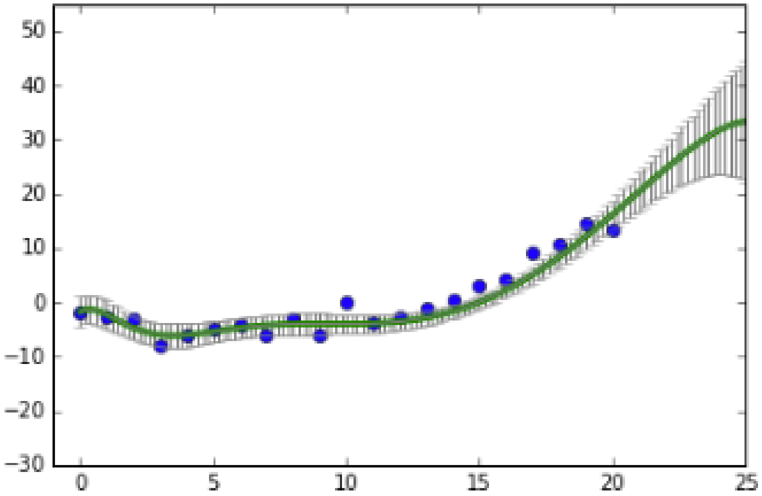


$$p(y^*|x^*, X, y) = \int_\theta p(y^*|x^*, \theta)p(\theta|X, y)d\theta$$

Retains information about the level of uncertainty around each prediction

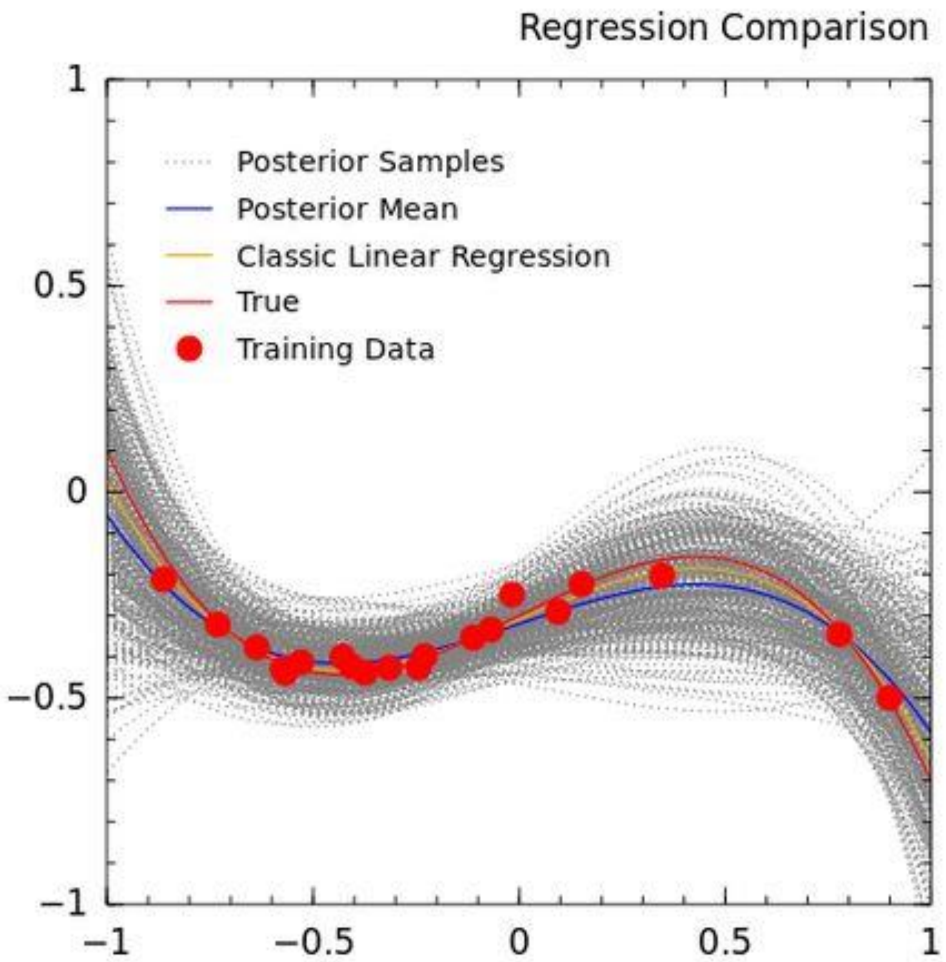
Bayesian Linear Regression !





$$p(y^*|x^*, X, y) = \int_\theta p(y^*|x^*, \theta)p(\theta|X, y)d\theta$$

Retains information about the level of uncertainty around each prediction


Regression Comparison
Posterior Samples
Posterior Mean
Classic Linear Regression
True
Training Data
1
0.5
0
−0.5
−1
−1
−0.5
0
0.5
1