Announcement

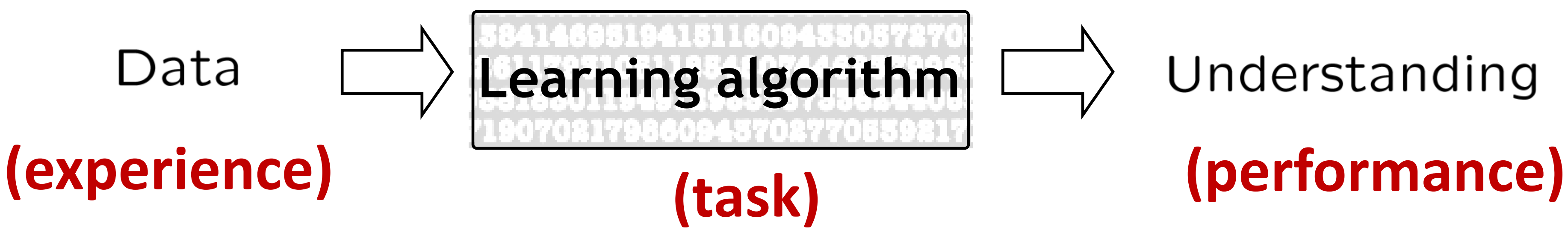
- HW 1 out TODAY – Watch your email

What is Machine Learning? (Formally)

What is Machine Learning?

Study of algorithms that

- improve their performance
- at some task
- with experience

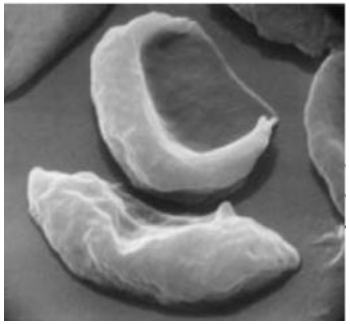


Supervised Learning Task

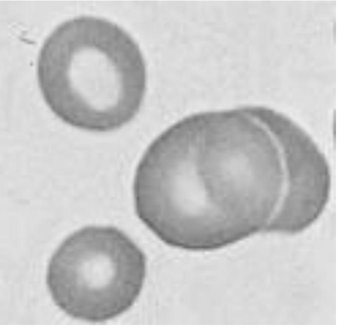
Task: Given $X \in \mathcal{X}$, predict $Y \in \mathcal{Y}$.

X - test data

$\equiv$ Construct **prediction rule** $f : \mathcal{X} \to \mathcal{Y}$



"Anemic cell (0)"

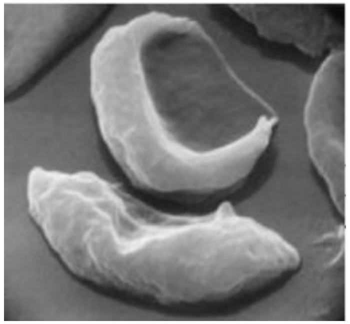


"Healthy cell (1)"

Performance Measures

Performance:

$\text{loss}(Y, f(X))$ - Measure of closeness between true label *Y* and prediction *f*(*X*)

X	*Y*	*f*(*X*)	$\text{loss}(Y, f(X))$
	"Anemic cell"	"Anemic cell"	0
		"Healthy cell"	1

$$\text{loss}(Y, f(X)) = 1_{\{f(X) \neq Y\}}$$ **0/1 loss**

Performance Measures

Performance:

$\text{loss}(Y, f(X))$ - Measure of closeness between true label *Y* and prediction *f*(*X*)

X	Share price, *Y*	*f*(*X*)	$\text{loss}(Y, f(X))$
Past performance, trade volume etc. as of Sept 8, 2010	"\$24.50"	"\$24.50"	0
		"\$26.00"	1?
		"\$26.10"	2?

$$\text{loss}(Y, f(X)) = (f(X) - Y)^2$$ **square loss**

Performance Measures

Performance:

$\text{loss}(Y, f(X))$ - Measure of closeness between true label *Y* and prediction *f*(*X*)

Don't just want label of one test data (cell image), but any cell image $X \in \mathcal{X}$

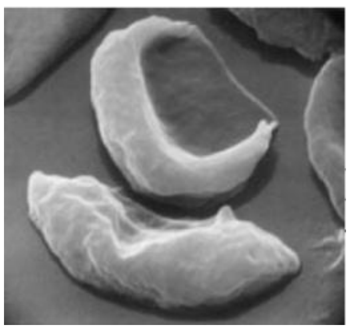
$$(X, Y) \sim P_{XY}$$

Given a cell image drawn randomly from the collection of all cell images, how well does the predictor perform on average?

$$\text{Risk } R(f) \equiv \mathbb{E}_{XY}\left[\text{loss}(Y, f(X))\right]$$

Performance Measures

Performance: Risk $R(f) \equiv \mathbb{E}_{XY}[\text{loss}(Y, f(X))]$



→ "Anemic cell"

→ Share Price "$ 24.50"

$\text{loss}(Y, f(X))$	Risk $R(f)$
$1_{\{f(X) \neq Y\}}$	$P(f(X) \neq Y)$
0/1 loss	**Probability of Error**
$(f(X) - Y)^2$	$\mathbb{E}[(f(X) - Y)^2]$
square loss	**Mean Square Error**

Bayes Optimal Rule

Ideal goal: Construct **prediction rule** $f^* : \mathcal{X} \to \mathcal{Y}$

$$f^* = \arg\min_f \mathbb{E}_{XY}\left[\text{loss}(Y, f(X))\right]$$

Bayes optimal rule

Best possible performance:

Bayes Risk $R(f^*) \leq R(f)$ for all f

BUT… Optimal rule is not computable - depends on unknown P_{XY} !

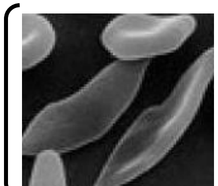
Experience - Training Data

Can't minimize risk since P_{XY} unknown!

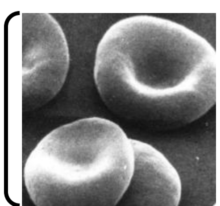
Training data (experience) provides a glimpse of P_{XY}

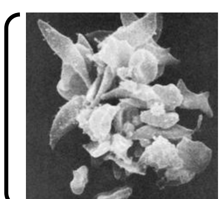
(observed) $\{(X_i, Y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} P_{XY}$ **(unknown)**

→ independent, identically distributed


, Anemic cell


, Healthy cell


, Healthy cell


, Anemic cell

Provided by expert, measuring device, some experiment, …

Supervised Learning

Task: Given $X \in \mathcal{X}$, predict $Y \in \mathcal{Y}$.

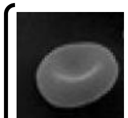
$\equiv$ Construct **prediction rule** $f : \mathcal{X} \to \mathcal{Y}$

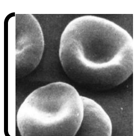
Performance: Risk $R(f) \equiv \mathbb{E}_{XY}\left[\text{loss}(Y, f(X))\right]$

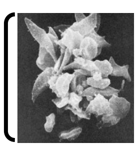
$$(X, Y) \sim P_{XY}$$

Experience: Training data $\{(X_i, Y_i)\}_{i=1}^{n} \overset{i.i.d.}{\sim} P_{XY}$ **(unknown)**

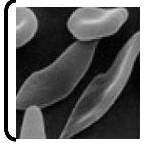

, Anemic cell

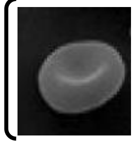

, Healthy cell

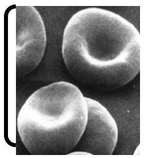

, Healthy cell

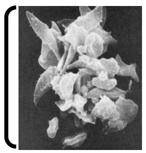

, Anemic cell

Machine Learning Algorithm


, Anemic cell

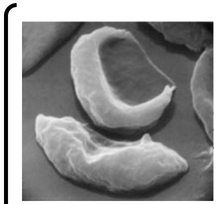

, Healthy cell


, Healthy cell


, Anemic cell

Training data $\{(X_i, Y_i)\}_{i=1}^{n}$ ⇒ **Learning algorithm** ⇒ $\widehat{f}_n$

$\widehat{f}_n$ is a mapping from $\mathcal{X} \to \mathcal{Y}$

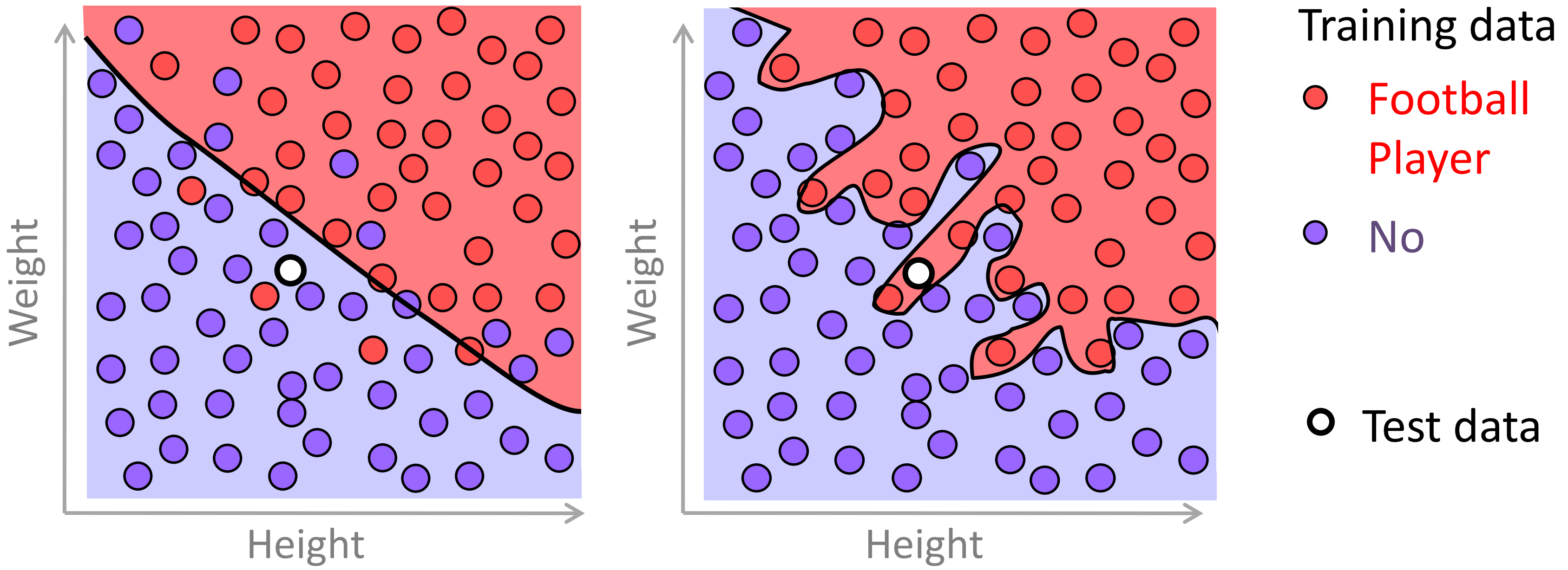
$\widehat{f}_n$ [] = "Anemic cell"

Test data X

Note: test data ≠ training data

Issues in ML

- A good machine learning algorithm
  - Does not **overfit** training data



  - **Generalizes** well to test data

More later …

Performance Revisited

Performance: (of a learning algorithm)

How well does the algorithm do on average

1. for a test cell image X drawn at random, and
2. for a set of training images and labels $D_n = \{(X_i, Y_i)\}_{i=1}^n$ drawn at random

Expected Risk (aka **Generalization Error**)

$$\mathbb{E}_{D_n}\left[R(\widehat{f}_n)\right] \equiv \mathbb{E}_{D_n}\left[\mathbb{E}_{XY}\left[\text{loss}(Y, \widehat{f}_n(X))\right]\right]$$

How to sense Generalization Error?

- Can't compute generalization error. How can we get a sense of how well algorithm is performing in practice?

- One approach -
  - Split available data into two sets $\{(X_i, Y_i)\}_{i=1}^n$ $\{(X'_i, Y'_i)\}_{i=1}^n$
  - Training Data – used for training the algorithm

$$\{(X_i, Y_i)\}_{i=1}^n \Rightarrow \text{Learning algorithm} \Rightarrow \widehat{f}_n$$

  - Test Data (a.k.a. Validation Data, Hold-out Data) – provides estimate of generalization error

Test Error = $\frac{1}{n}\sum_{i=1}^{n}\left[\text{loss}(Y'_i, \widehat{f}_n(X'_i))\right]$

Why not use Training Error?

Supervised vs. Unsupervised Learning

Supervised Learning – Learning with a teacher

$\{(X_i, Y_i)\}_{i=1}^n$ → **Learning algorithm** → $\widehat{f}_n$

Documents, topics

Mapping between Documents and topics

Unsupervised Learning – Learning without a teacher

$\{X_i\}_{i=1}^n$ → **Learning algorithm** → $\widehat{f}_n$

Documents

Model for word distribution OR Clustering of similar documents

Lets get to
some learning algorithms!

Learning Distributions (Parametric Approach)

Aarti Singh

Machine Learning 10-701/15-781
Sept 13, 2010

Your first consulting job

- A billionaire from the suburbs of Seattle asks you a question:
  - He says: I have a coin, if I flip it, what's the probability it will fall with the head up?
  - You say: Please flip it a few times:

  - You say: The probability is: **3/5**
  - **He says: Why???**
  - You say: Because…

Bernoulli distribution

Data, D = 

- P(Heads) = θ, P(Tails) = 1-θ
- Flips are **i.i.d.**:
  - **Independent** events
  - **Identically distributed** according to Bernoulli distribution

<u>Choose θ that maximizes the probability of observed data</u>

Maximum Likelihood Estimation

Choose θ that maximizes the probability of observed data

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

MLE of probability of head:

$$\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$ **= 3/5**

"Frequency of heads"

How many flips do I need?

$$\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

- Billionaire says: I flipped 3 heads and 2 tails.
- You say: θ = 3/5, I can prove it!
- He says: What if I flipped 30 heads and 20 tails?
- You say: Same answer, I can prove it!
- **He says: What's better?**
- You say: Hmm… The more the merrier???
- He says: Is this why I am paying you the big bucks???

Simple bound (Hoeffding's inequality)

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

- Let θ^* be the true parameter, for any $\varepsilon > 0$:

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

PAC Learning

- PAC: Probably Approximate Correct
- Billionaire says: I want to know the coin parameter θ, within ε = 0.1, with probability at least 1-δ = 0.95. How many flips?

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

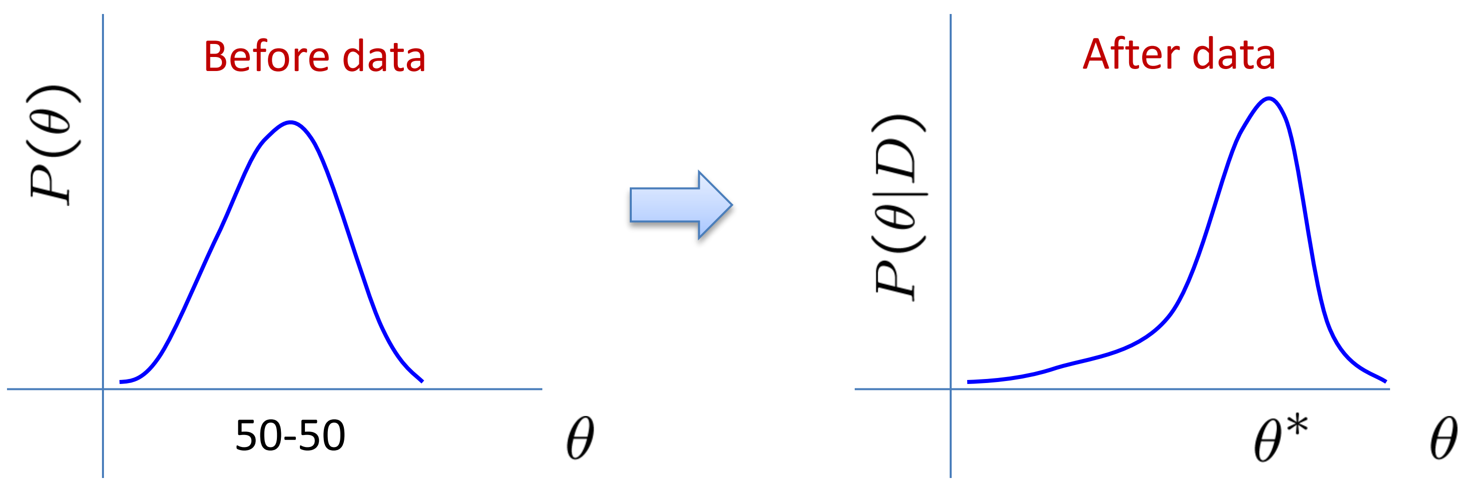
Sample complexity

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

What about prior knowledge?

- Billionaire says: Wait, I know that the coin is "close" to 50-50. What can you do for me now?
- **You say: I can learn it the Bayesian way…**

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



Bayesian Learning

- Use Bayes rule:

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

- Or equivalently:

$$P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta)$$

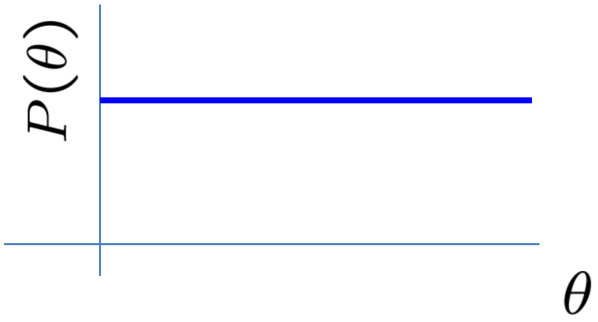
posterior likelihood prior

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**

Prior distribution

- What about 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

Conjugate Prior

- P(θ) and P(θ|D) have the same form

Eg. 1 Coin flip problem

Likelihood is ~ Binomial

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

If prior is Beta distribution,

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

Then posterior is Beta distribution

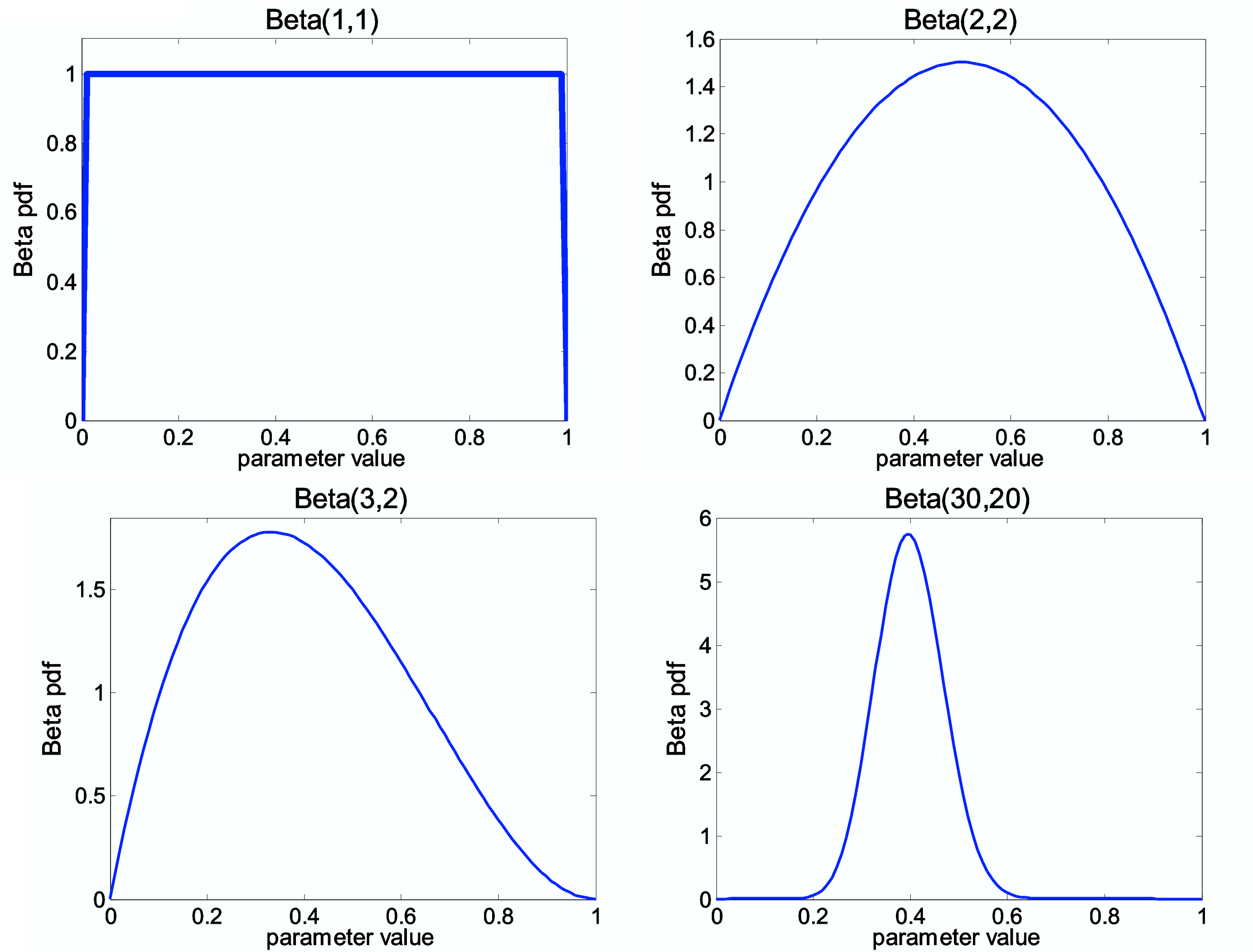
$$P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

For Binomial, conjugate prior is Beta distribution.

Beta distribution

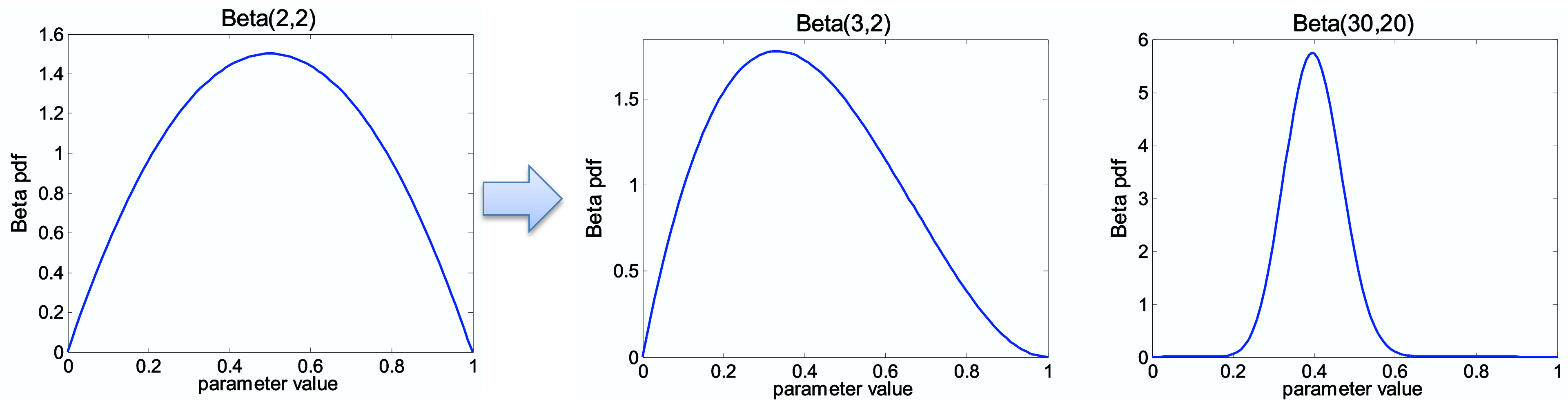
$Beta(\beta_H, \beta_T)$

More concentrated as values of β_H, β_T 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”

Conjugate Prior

- $P(\theta)$ and $P(\theta|D)$ have the same form

Eg. 2 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.

Maximum A Posteriori Estimation

Choose θ that maximizes a posterior probability

$$\begin{aligned}\widehat{\theta}_{MAP} &= \arg\max_{\theta} \quad P(\theta \mid D) \\ &= \arg\max_{\theta} \quad P(D \mid \theta) P(\theta)\end{aligned}$$

MAP estimate of probability of head:

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

$$\widehat{\theta}_{MAP} = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}$$

Mode of Beta distribution