Probability and Naive Bayes

Machine Learning
CSx824/ECEx242
Bert Huang
Virginia Tech

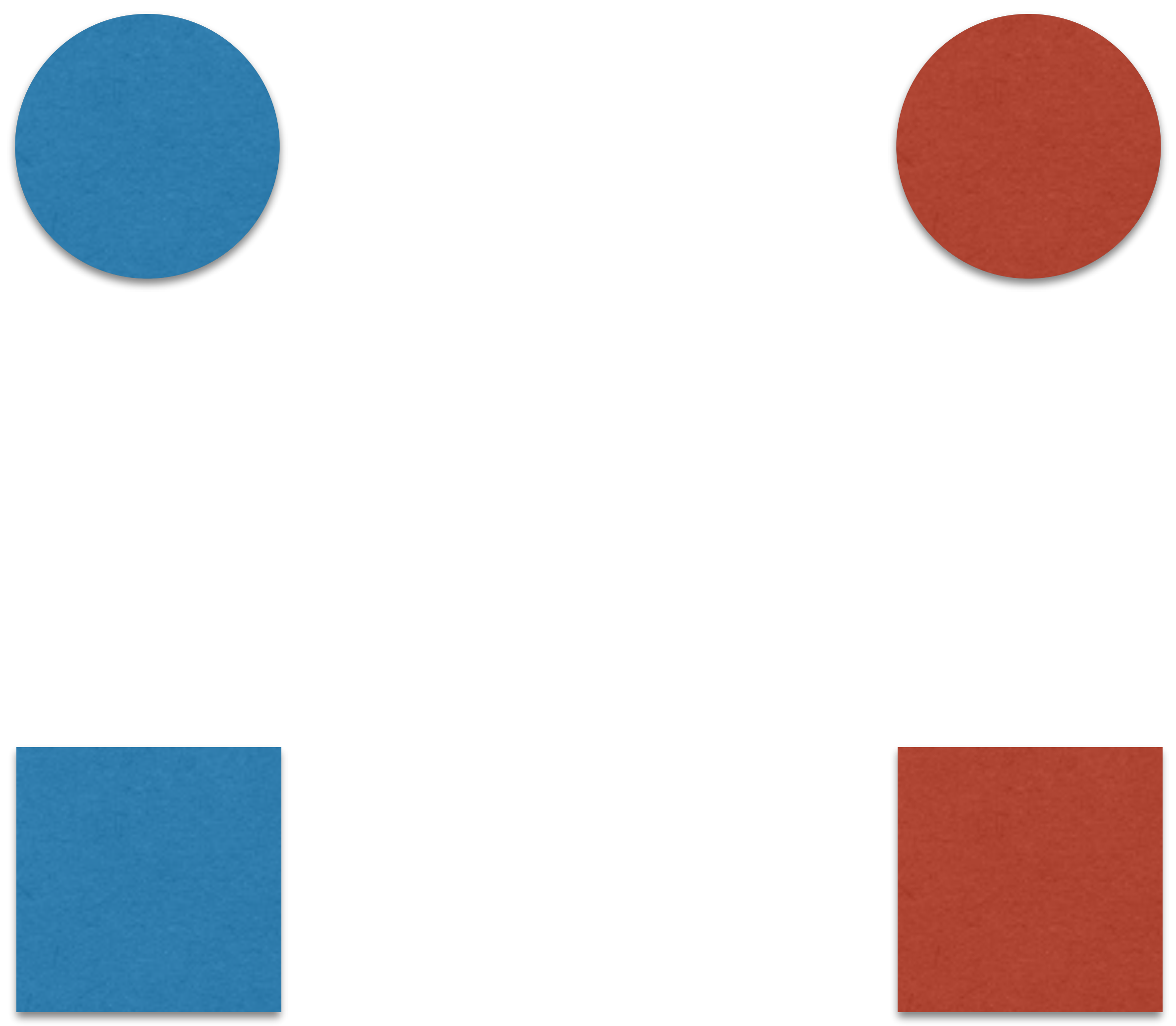
Outline

- Probabilistic identities
- Independence and conditional independence
- Naive Bayes
- Log tricks

Probability Identities

- Random variables in caps (**A**)
  - values in lowercase: **A = a** or just **a** for shorthand
- $P(a \mid b) = P(a, b) / P(b)$ — conditional probability
- $P(a, b) = P(a \mid b) P(b)$ — joint probability
- $P(b \mid a) = P(a \mid b) P(b) / P(a)$

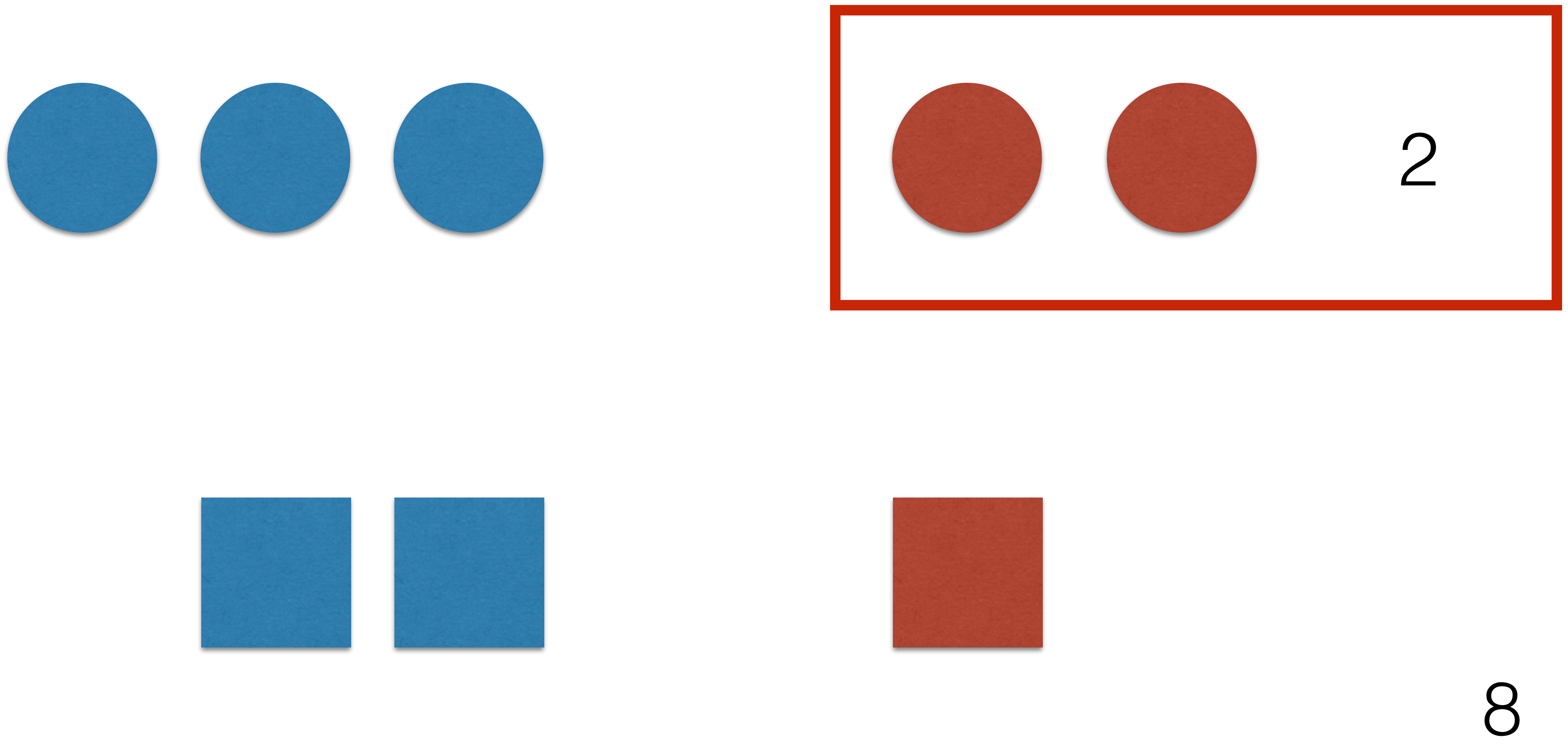
Probability via Counting



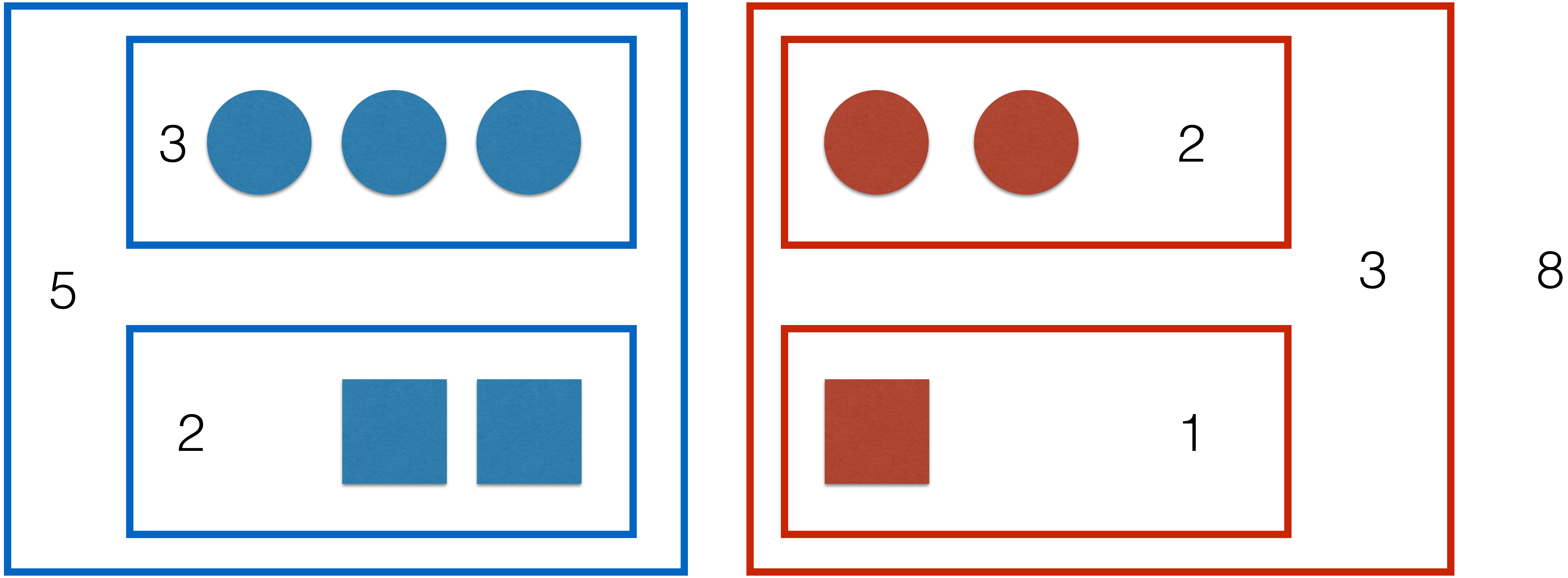
Probability via Counting

P(circle, red)

=2/8 = 0.25



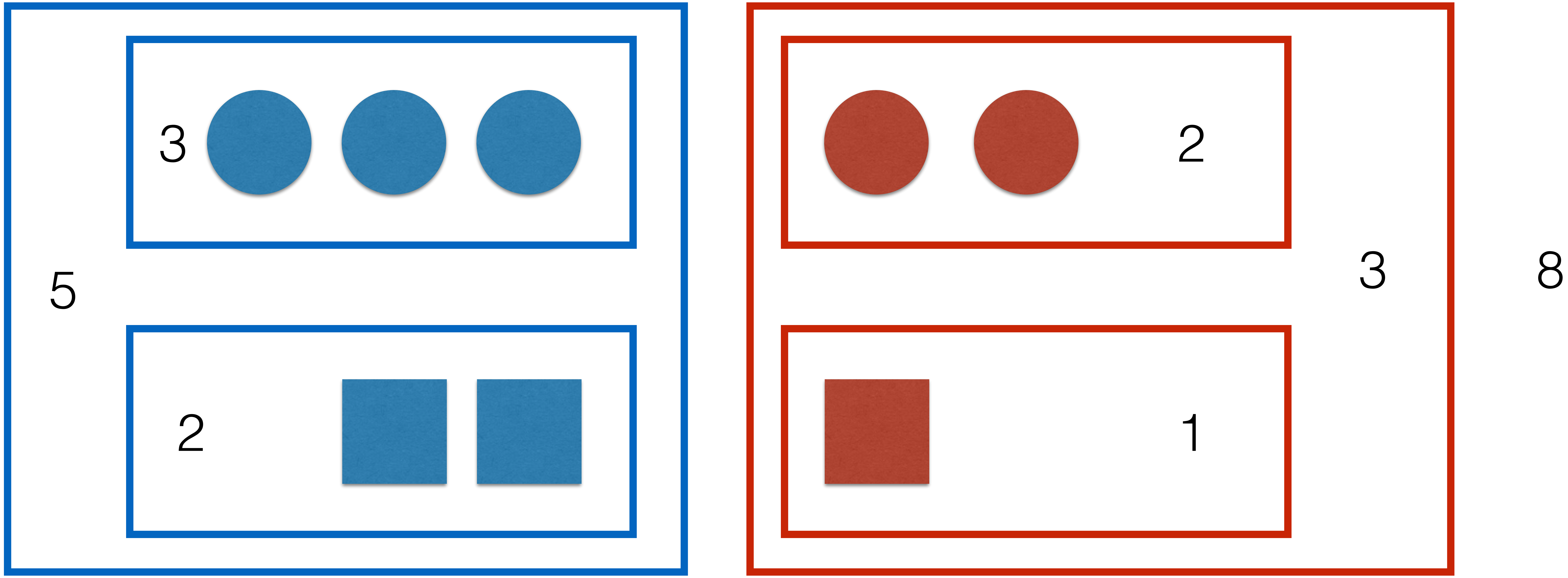
Probability via Counting



P(circle | red) = P(circle, red) / P(red)

2/3 2/8 3/8

Probability via Counting



P(circle | red) P(red) = P(circle, red)

2/3 3/8 2/8

Probability Identities

- Random variables in caps (**A**)
  - values in lowercase: **A = a** or just **a** for shorthand
- $P(a \mid b) = P(a, b) / P(b)$
- $P(a, b) = P(a \mid b) P(b)$
- $P(b \mid a) = P(a \mid b) P(b) / P(a)$

Bayes Rule

- $P(b \mid a)$
- $P(b \mid a) = P(a, b) / P(a)$
- $P(b \mid a) = P(a \mid b) P(b) / P(a)$

Classification

- $x \in \{0,1\}^d$, $y \in \{0,1\}$
- $f(x) \in \{0,1\}$
- Accuracy: $E[f(x) = y]$
- **Bayes optimal classifier:** $f(x) = \arg\max_y p(y \mid x)$
  - Seems natural, but why is this optimal?

Back-of-Envelope for Bayes Optimal

- For each unique **x**, **p(y | x)** is a coin flip
- Assume we need **n** samples to accurately estimate a coin flip
- How many unique **x**'s?
    - $|\{0,1\}^d| = 2^d$
- Need $\mathbf{n2^d}$ samples

$n = 100$

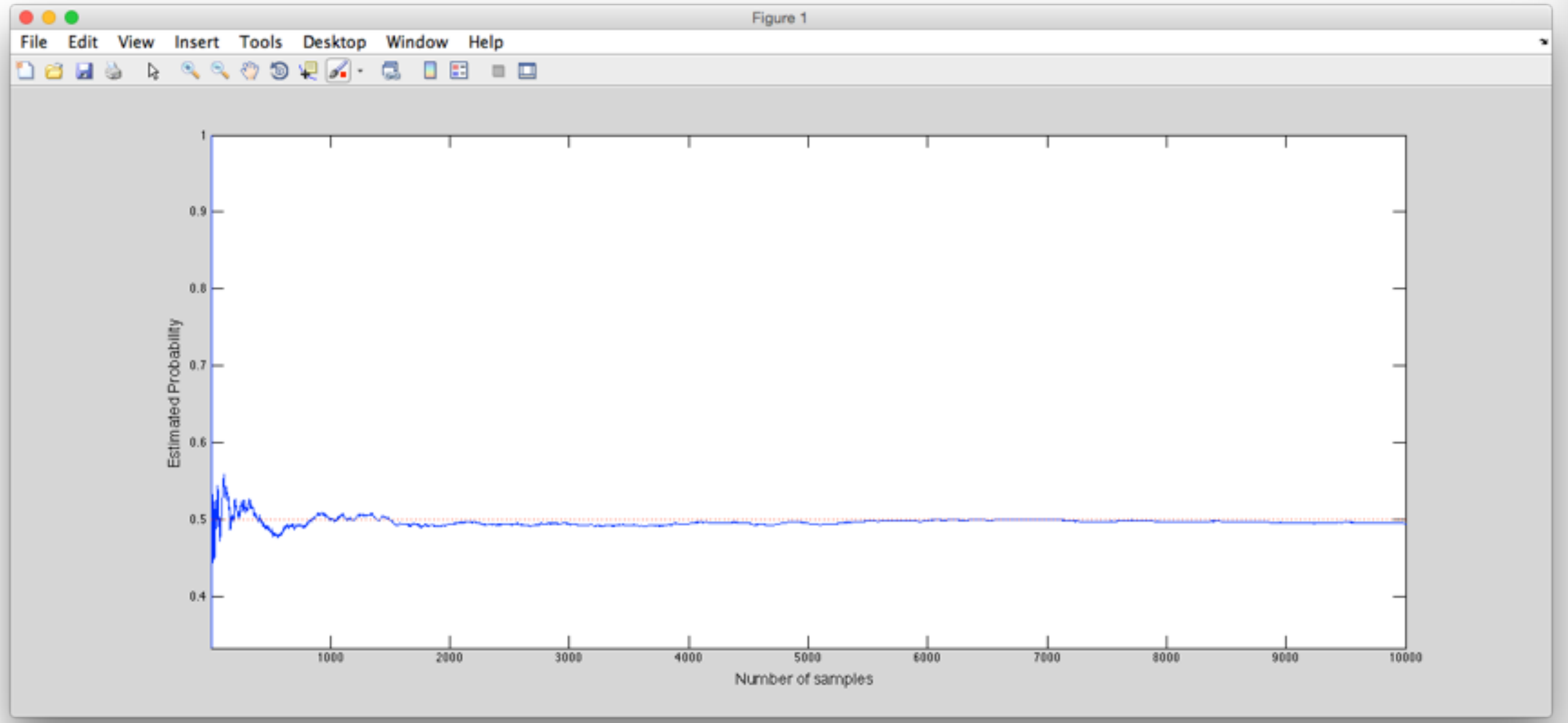
$d = 100$

Need 1.2676506×10^{32} samples

1.2676506 x 10,000,000,000,000,000,000,000,000,000,000

How Many Samples?

- Concentration bounds



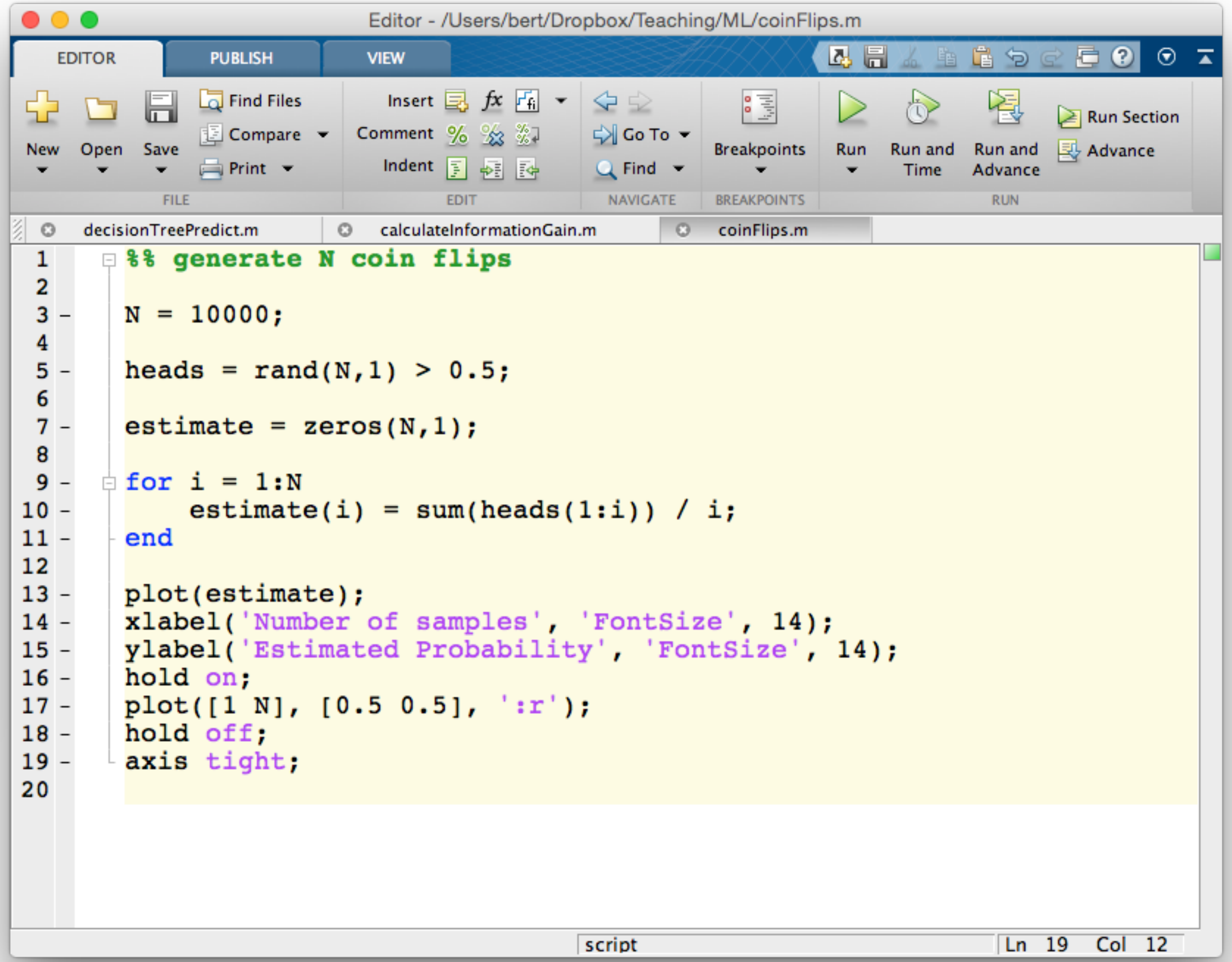


Figure 1
File Edit View Insert Tools Desktop Window Help
Estimated Probability
Number of samples
1
0.9
0.8
0.7
0.6
0.5
0.4
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000

Independence

- **A** and **B** are independent iff $p(A, B) = p(A)\, p(B)$
- **A** and **B** are *conditionally* independent given **C** iff
  $p(A, B \mid C) = p(A \mid C)\, p(B \mid C)$
- $|\, p(A, B) \,| = |\, A \,| \times |\, B \,|$ $\qquad |\, p(A) \,| = |\, A \,|$ $\qquad |\, p(B) \,| = |\, B \,|$

Naive Bayes

- Assume dimensions of **x** are conditionally independent given **y**
  - Bag of words: p("virginia", "tech" | y) = p("virginia" | y) p("tech" | y)
- $f(x) = \arg\max_y p(y \mid x)$
- $= \arg\max_y p(x \mid y)\, p(y) / p(x)$
- $= \arg\max_y p(x \mid y)\, p(y)$
- $= \arg\max_y p(y) \prod_j p(x_j \mid y)$

Bernoulli Maximum Likelihood

- $p(y) \prod_j p(x_j \mid y)$
- $p(Y = y) \leftarrow$ (# examples where $Y = y$) / (# examples)
- $p(X_j = x_j \mid y) \leftarrow$ (# ex. where $Y = y$ and $X_j = x_j$) / (# ex. where $Y=y$)
- Learning by counting!

Breaking Maximum Likelihood

- Happy: “Great!”
  Happy: “Had a great day”
  Sad: “:-(Bad day”
  Sad: “:-(”

great	had	a	bad	day	:-(
1	0	0	0	0	0
1	1	1	0	1	0
0	0	0	1	1	1
0	0	0	0	0	1

- ???: “Had a bad day :-(”

- $p(y) = 0.5$

	great	had	a	bad	day	:-(
Happy	1.0	0.5	0.5	0.0	0.5	0
Sad	0.0	0.0	0.0	0.5	0.5	1.0

- $p(\text{happy} \mid \ldots) \propto 0.5 \times 0.0 \ldots$

  $p(\text{sad} \mid \ldots) \propto 0.5 \times 1.0 \times 0.0 \times \ldots$

Breaking Maximum Likelihood

great	had	a	bad	day	:-(
1	0	0	0	0	0
1	1	1	0	1	0
0	0	0	1	1	1
0	0	0	0	0	1

- Happy: "Great!"
  Happy: "Had a great day"
  Sad: ":-(Bad day"
  Sad: ":-("
- ???: "Had a bad day :-("
- p(y) = 0.5

	great	had	a	bad	day	:-(
Happy	1.0	0.5	0.5	0.0	0.5	0
Sad	0.0	0.0	0.0	0.5	0.5	1.0

- p(happy | …) ∝ 0.5 x 0.0 …

  p(sad | …) ∝ 0.5 x 1.0 x 0.0 x …

Fixing Maximum Likelihood

- p(Z = z) ← (# examples where Y = y + **α**) / (# examples + 2**α**)
    - E.g., **α** = 1
- **α** vanishes as # of examples grows toward infinity
- When # is small, **α** prevents 1.0 or 0.0 estimates

Breaking Maximum Likelihood

- Happy: “Great!”
  Happy: “Had a great day”
  Sad: “:-(Bad day”
  Sad: “:-(”

great	had	a	bad	day	:-(
1	0	0	0	0	0
1	1	1	0	1	0
0	0	0	1	1	1
0	0	0	0	0	1

$$\frac{2 + 1}{2 + 2\,(1)} = 3/4$$

- ???: “Had a bad day :-(”

- p(y) = 0.5

	great	had	a	bad	day	:-(
Happy	0.75	0.5	0.5	0.25	0.5	0.25
Sad	0.25	0.25	0.25	0.5	0.5	0.75

- $p(\text{happy} \mid \ldots) \propto 0.5 \times 0.25 \times 0.5 \times 0.5 \times 0.25 \times 0.5 \times 0.75 = 0.0029$

  $p(\text{sad} \mid \ldots) \propto 0.5 \times 0.75 \times 0.25 \times 0.25 \times 0.5 \times 0.5 \times 0.75 = 0.0044$

Maximum a Posteriori

- Bernoulli: $p(Z \mid \theta) = \theta^{Z} (1 - \theta)^{(1-Z)}$
- Maximum likelihood: $\theta \leftarrow \arg\max_{\theta'} p(Z \mid \theta')$
- Maximum a posteriori = maximize posterior: $\theta \leftarrow \arg\max_{\theta'} p(\theta' \mid Z)$
- $p(\theta' \mid Z) = p(Z \mid \theta')\, p(\theta') / p(Z)$
- MAP: $\theta \leftarrow \arg\max_{\theta'} p(Z \mid \theta')\, p(\theta')$
- Previous trick equiv. to setting $p(\theta')$ to a Beta distribution

Continuous Data

- Conditional feature independence with continuous data?
- E.g., use normal distribution for $p(x_j \mid y)$
  - $p(y)$ is the same as before
  - $p(x_j \mid y)$ is the MLE for univariate normal

Log Tricks

- Each $p(x_j \mid y)$ is in [0,1]
- Multiplying **d** of them quickly goes to numerical zero
  - E.g., $0.9^{256} = 1.932334983E\text{-}12$
- Instead, use log probabilities: $\log \prod_j p(x_j \mid y) = \sum_j \log p(x_j \mid y)$
  - E.g., $\log 0.9^{256} = 256 \log 0.9 = -11.71$

Summary

- Probabilistic identities
- Independence and conditional independence
- Naive Bayes
- Log tricks