More Naïve Bayes

Previously on CSCI 4622

Probabilistic Classification: $\hat{y} = \arg\max_c p(y = c \mid \mathbf{x})$

Generative Models estimate these probabilities from **Bayes Rule:** $p(y \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid y)p(y)}{p(\mathbf{x})}$

Throw out denominator and compare **numerators:** $p(\text{SPAM} \mid \mathbf{x}) \propto p(\mathbf{x} \mid \text{SPAM})p(\text{SPAM})$

Make Naïve **Conditional Independence assumption:**

$$p(\mathbf{x} = [\text{buy, viagra}] \mid SPAM) = p(\text{buy} \mid \text{SPAM}) \cdot p(\text{viagra} \mid \text{SPAM})$$

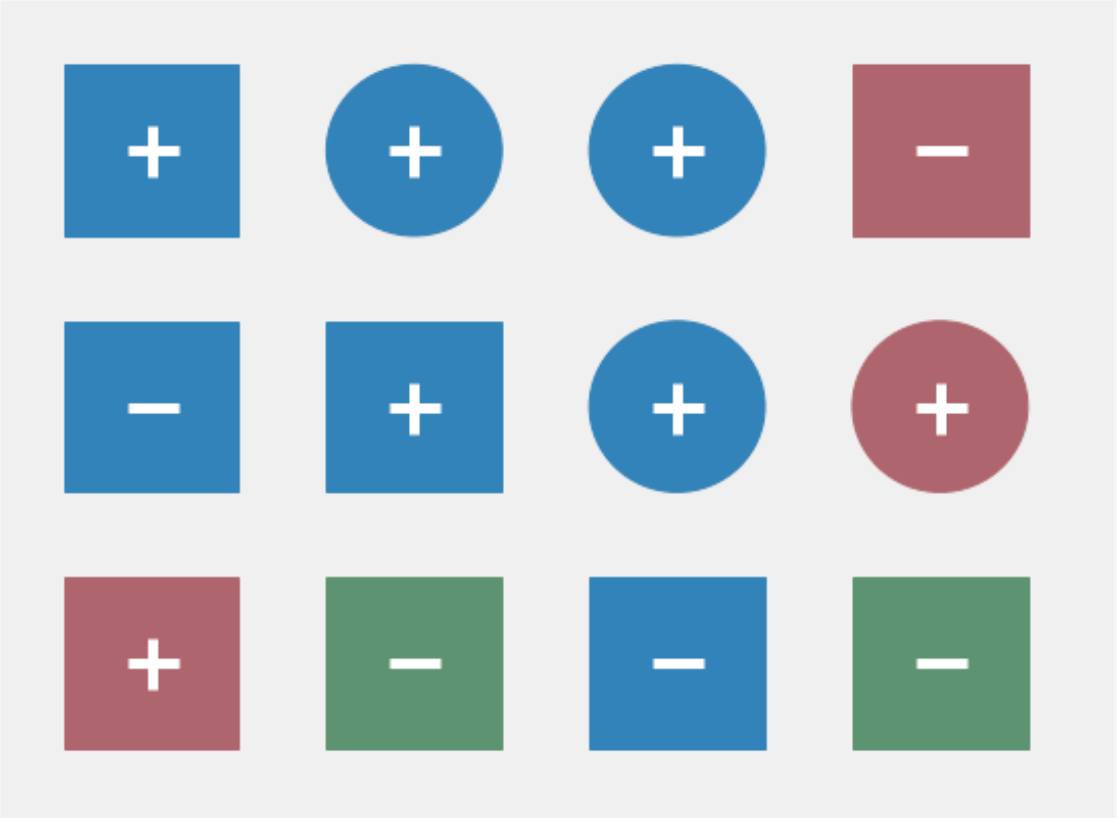
Estimate likelihoods and priors from text, multiply to get scores, and make predictions

$$\hat{p}(\text{term} \mid \text{Class}) = \frac{\#\text{ instances of term in Class} + 1}{\#\text{ total } \mathbf{words} \text{ in Class} + |V|} \qquad \hat{p}(\text{Class}) = \frac{\#\text{ emails from Class} + 1}{\#\text{ total emails in training data} + |C|}$$

Slightly Different Warm-Up Exercise

Example: Consider the following binary labeled training set where labels are + and −

Question: What are the features?



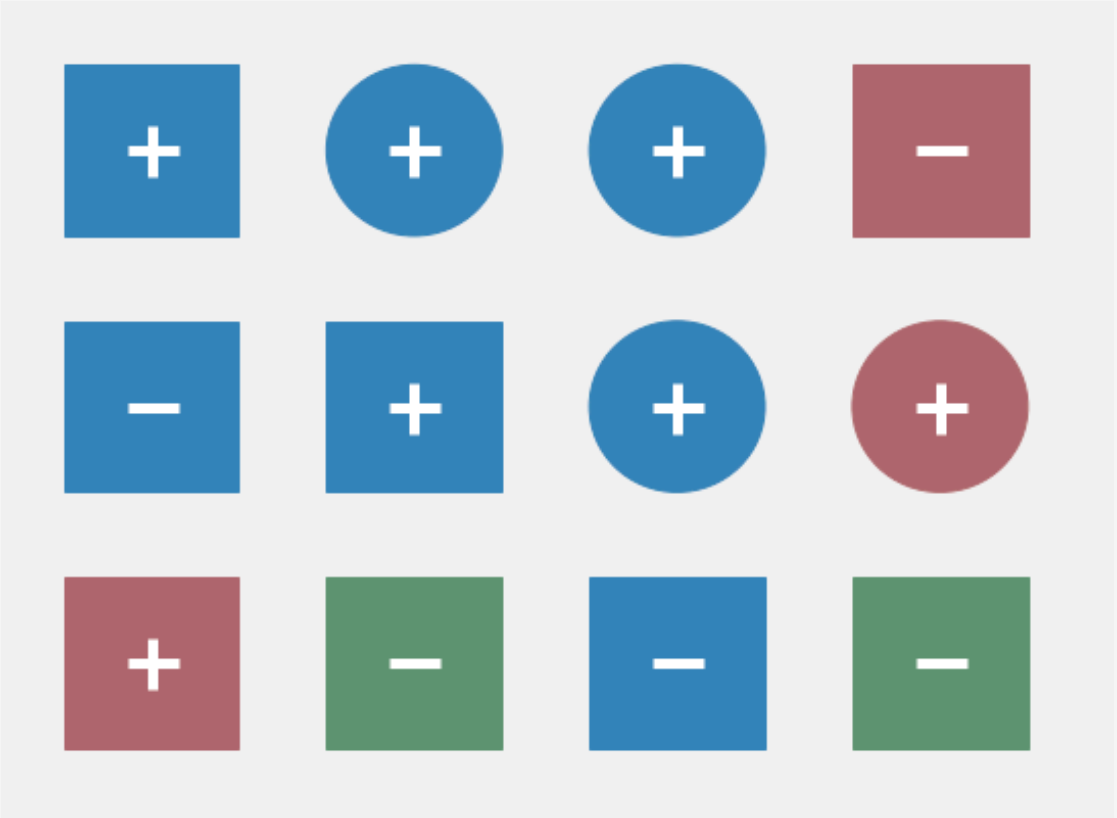
Slightly Different Warm-Up Exercise

Example: Consider the following binary labeled training set where labels are + and −

Question: What are the features?

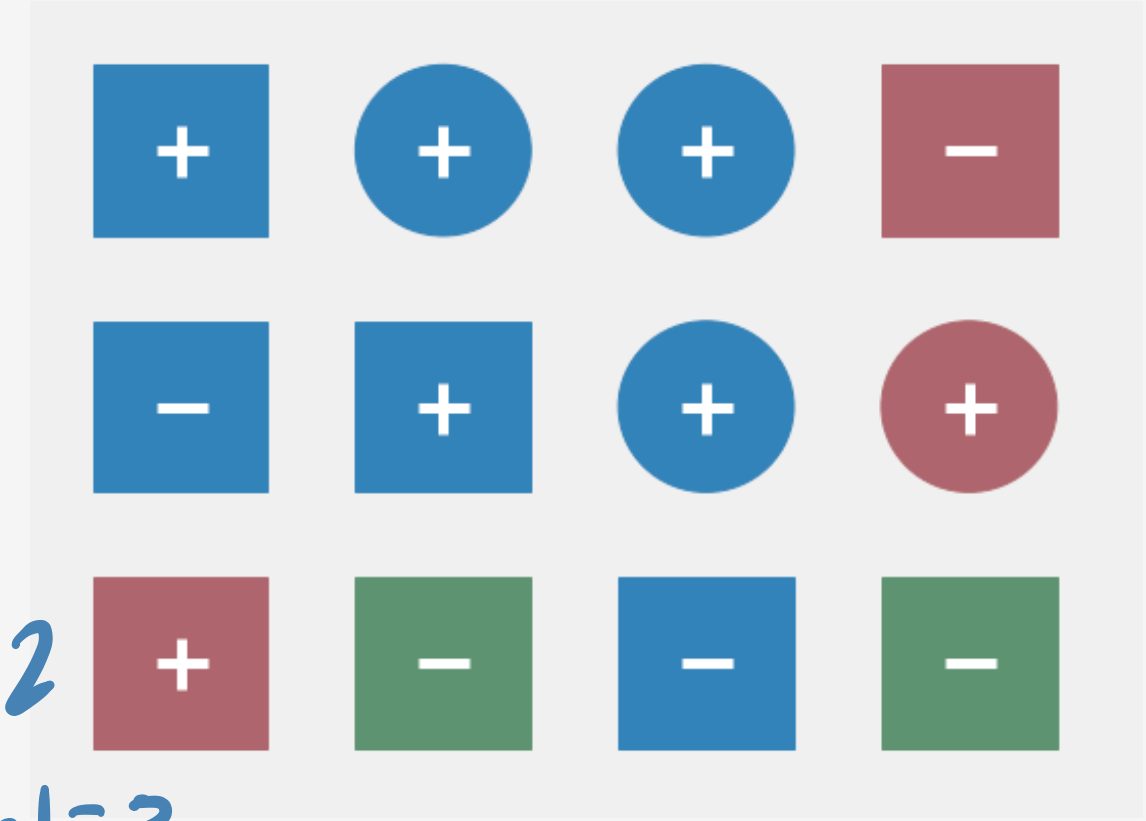
Answer: Shapes and Colors

Question: What are the feature vocabs?



Slightly Different Warm-Up Exercise

Example: Consider the following binary labeled training set where labels are + and −



Question: What are the features?

Answer: Shapes and Colors

Question: What are the feature vocabs?

Answer:

$x_s \in \{\text{square, circle}\} = V_s \quad |V_s| = 2$

$x_c \in \{\text{red, blue, green}\} = V_c \quad |V_c| = 3$

Slightly Different Warm-Up Exercise

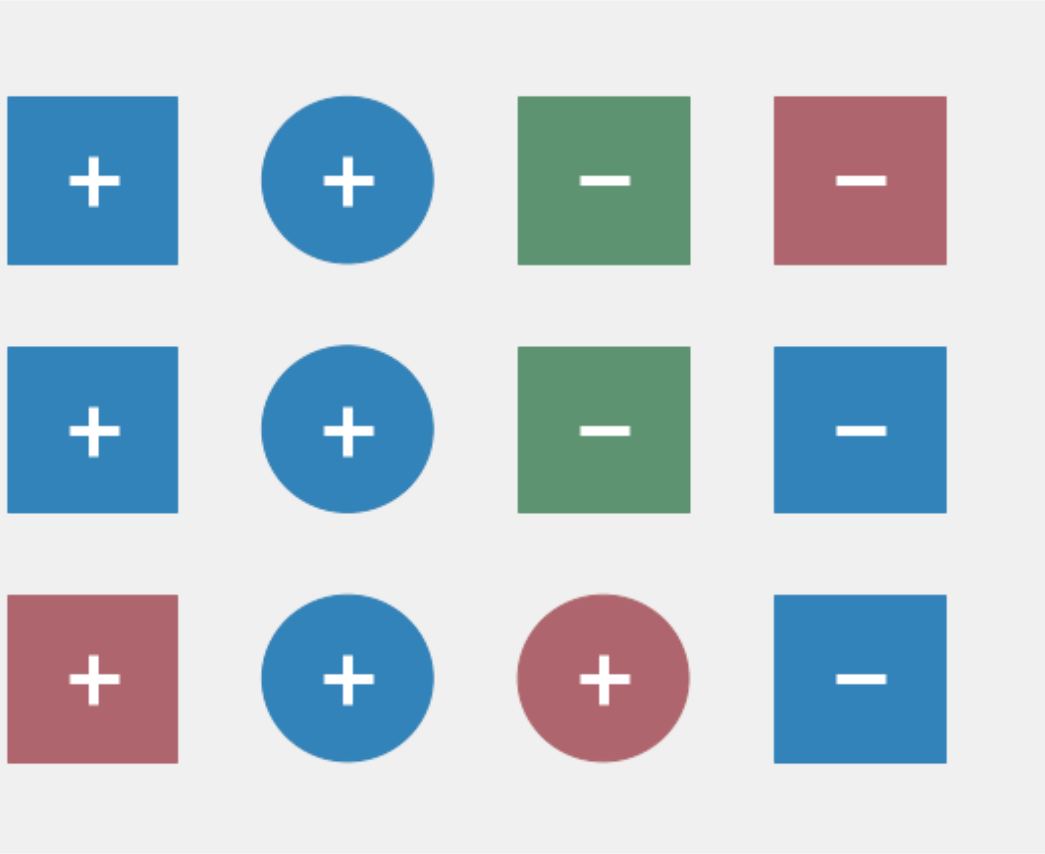
Question: How would we classify  ?

Need to compute $p(+ \mid \mathbf{x} = [\text{blue, square}]) \propto$

$$\hat{p}(\text{blue} \mid +) \cdot \hat{p}(\text{square} \mid +) \cdot \hat{p}(+) =$$

$$\frac{5}{7} \cdot \frac{3}{7} \cdot \frac{7}{12}$$

$$= 0.179$$



Slightly Different Warm-Up Exercise

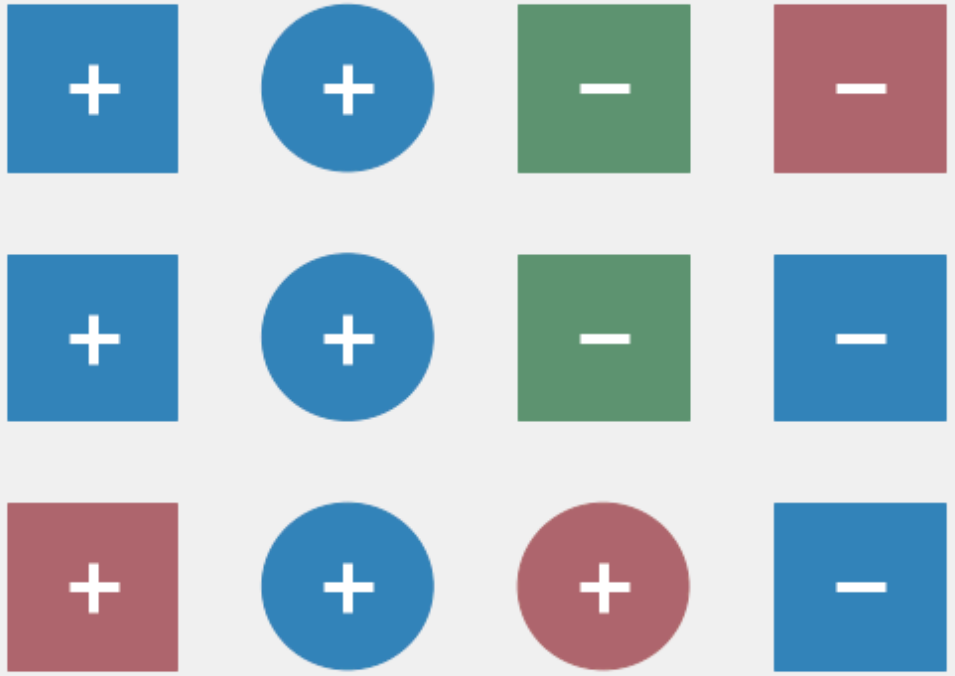
Question: How would we classify  ?

Need to compute $p(- \mid \mathbf{x} = [\text{blue, square}]) \propto$

$$\hat{p}(\text{blue} \mid -) \cdot \hat{p}(\text{square} \mid -) \cdot \hat{p}(-) =$$

$$\frac{2}{5} \cdot \frac{5}{5} \cdot \frac{5}{12} =$$

$$0.167$$



Slightly Different Warm-Up Exercise

Question: How would we classify  ?

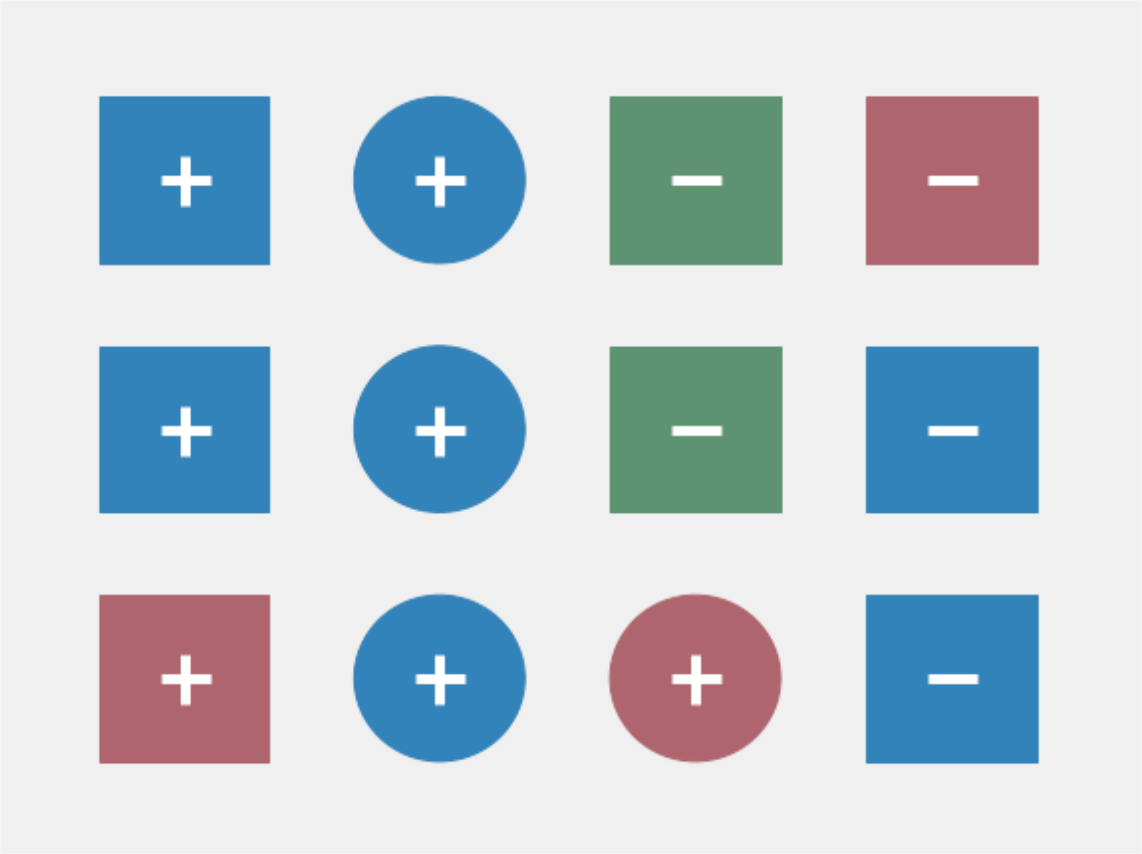
So we have:

$$p(+ \mid \mathbf{x} = [\text{blue, square}]) \propto 0.179$$

$$>$$

$$p(- \mid \mathbf{x} = [\text{blue, square}]) \propto 0.167$$

So we predict POSITIVE



Slightly Different Warm-Up Exercise

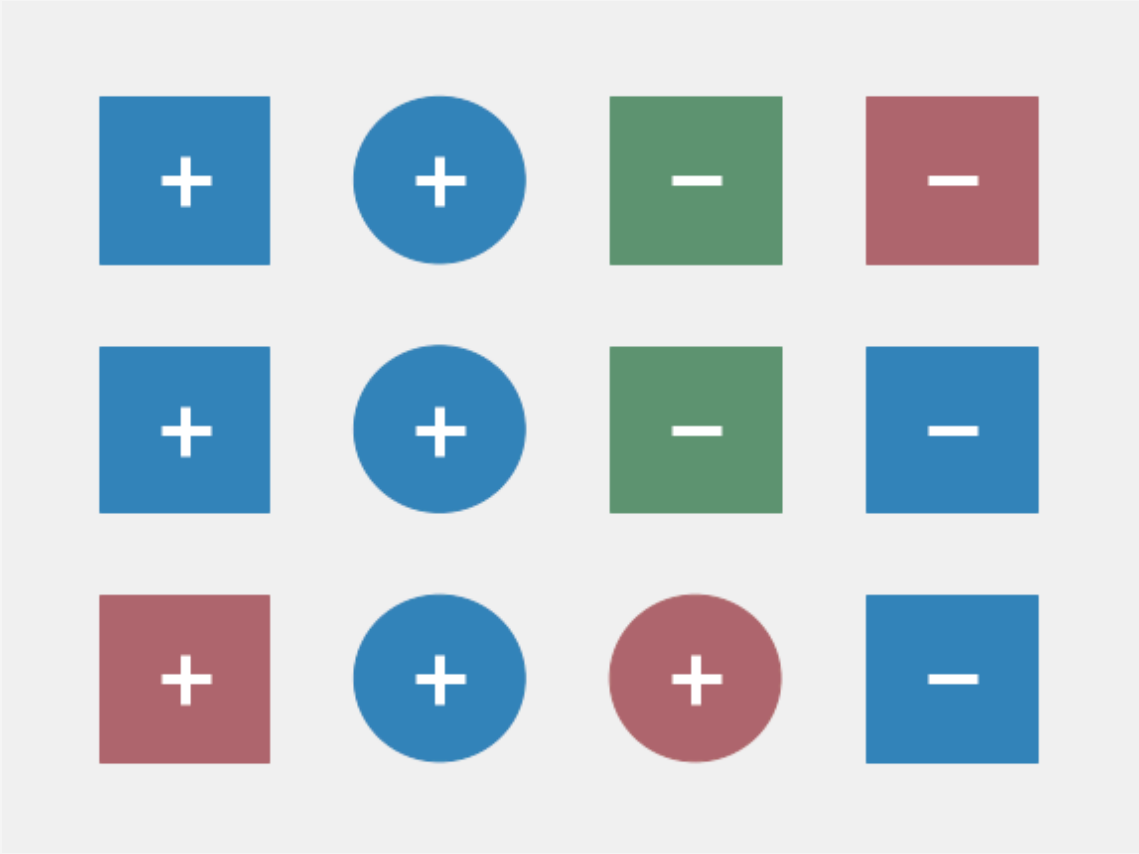
Question: How would we classify  ?

So we have:

$$p(+ \mid \mathbf{x} = [\text{blue, square}]) \propto 0.179$$

$$p(- \mid \mathbf{x} = [\text{blue, square}]) \propto 0.167$$

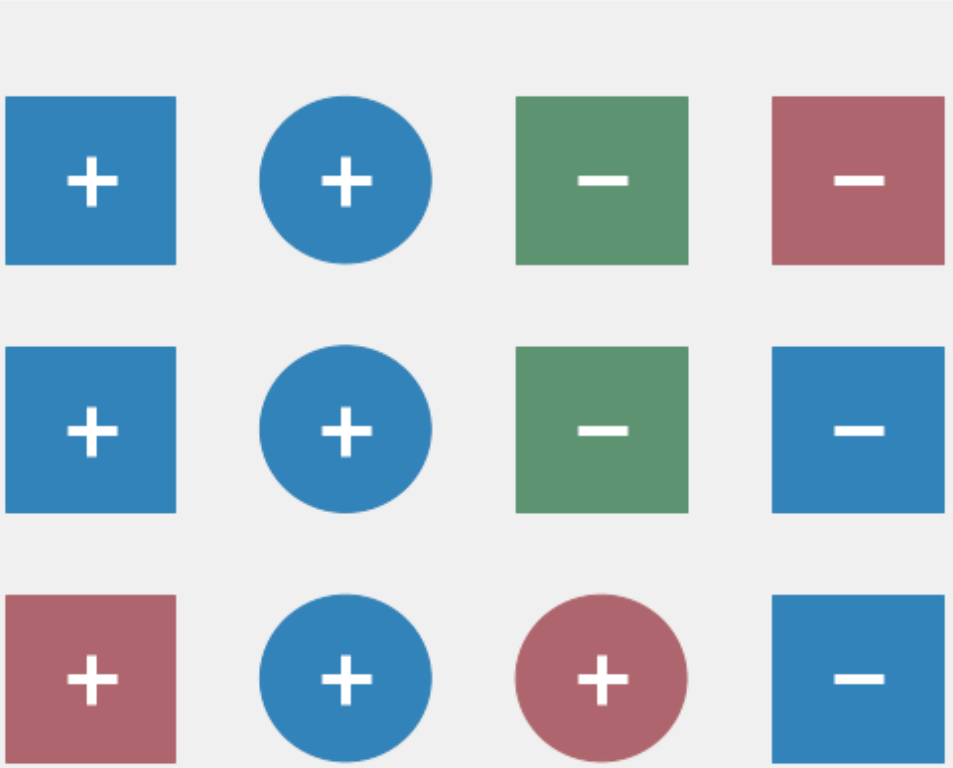
So we predict positive .



Question: Is this what you would have predicted for  ?

Slightly Different Warm-Up Exercise

Question: How would we classify  ?



So we have:

$$p(+ \mid \mathbf{x} = [\text{blue, square}]) \propto 0.179$$

$$p(- \mid \mathbf{x} = [\text{blue, square}]) \propto 0.167$$

So we predict positive .

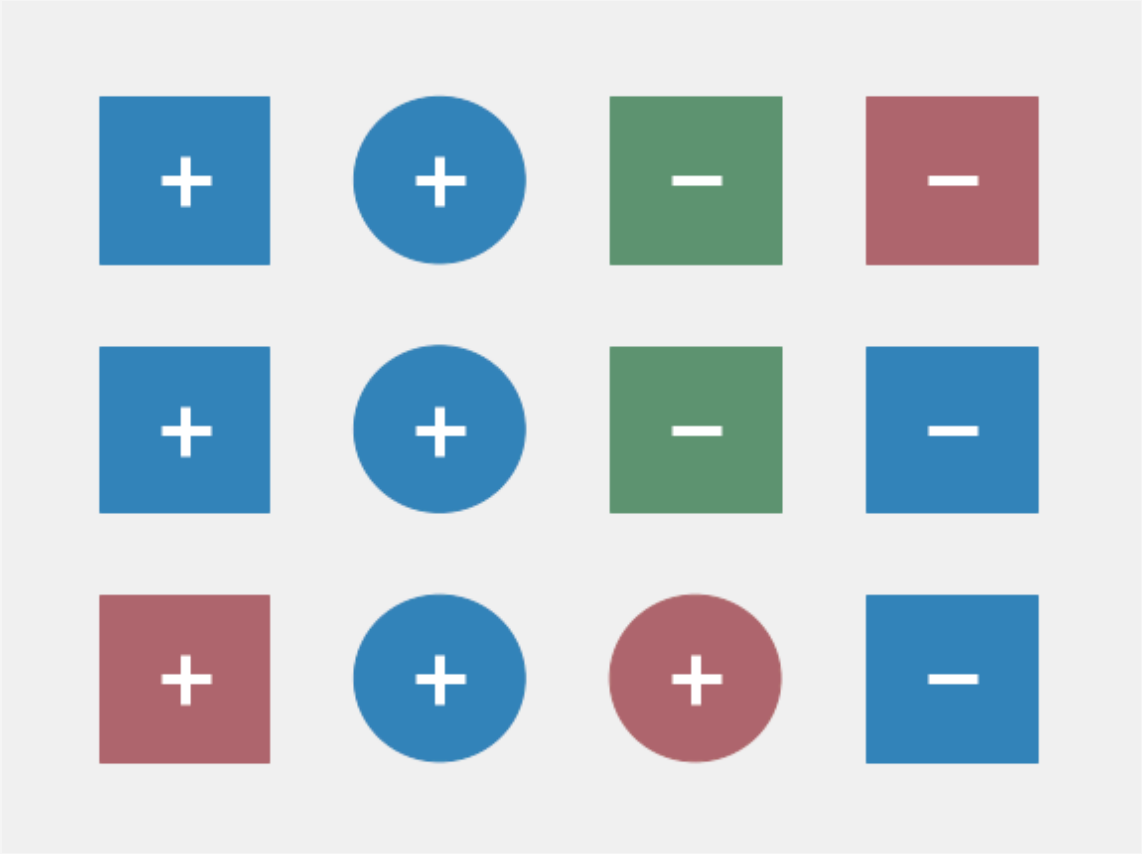
Question: Is this what you would have predicted for  ?

Observation: If we had only looked at the class-likelihood, it would have indicated **negative**, but there are several more positive examples, so the **prior** pushes the prediction to **positive**

Slightly Different Warm-Up Exercise

Question: How would we classify  ?

First: Do you see any difficulties with  ?



Slightly Different Warm-Up Exercise

Question: How would we classify  ?

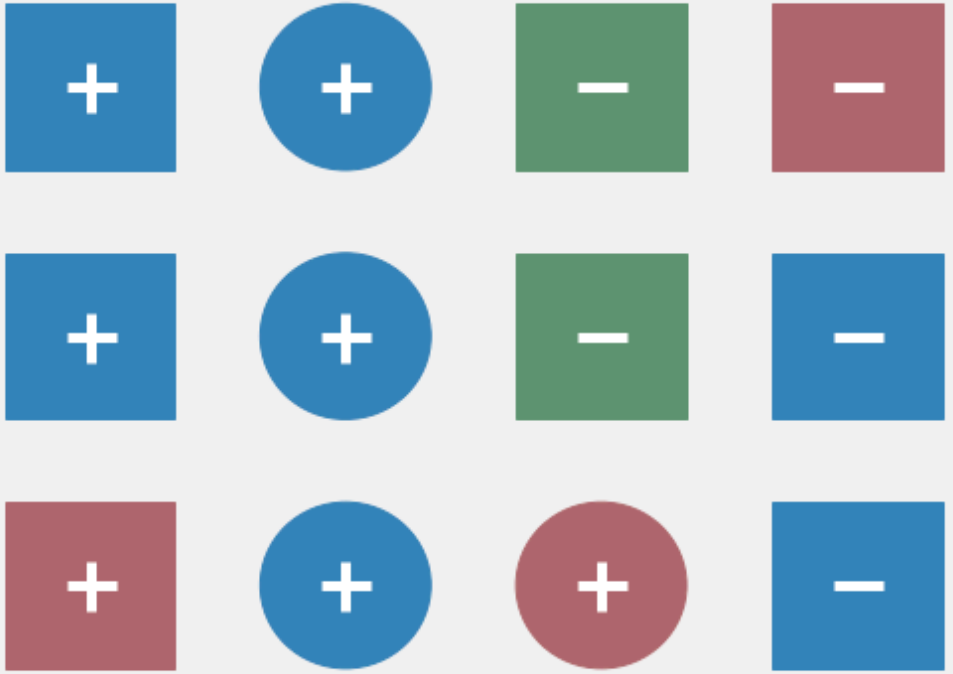
Need to compute $p(+ \mid \mathbf{x} = [\text{green}, \text{circle}]) \propto$

$$\hat{p}(\text{green} \mid +) \cdot \hat{p}(\text{circle} \mid +) \cdot \hat{p}(+) =$$

$$\frac{0+1}{7+3} \cdot \frac{4+1}{7+2} \cdot \frac{7+1}{12+2}$$

ADD-ONE

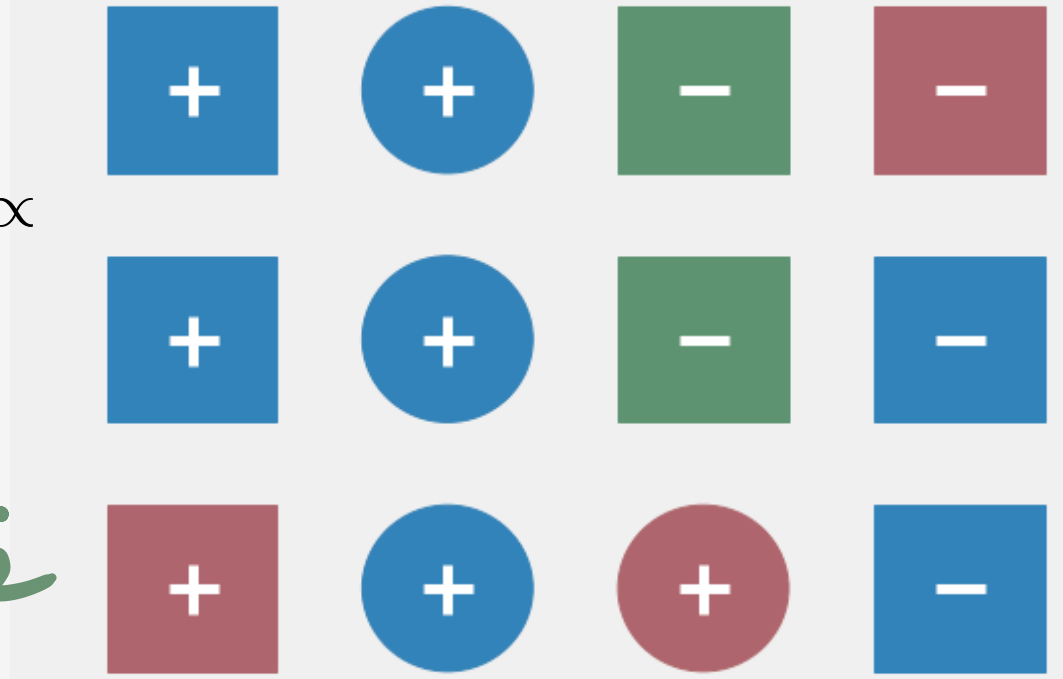
THREE COLORS — TWO SHAPES



$$= \frac{1}{10} \cdot \frac{5}{9} \cdot \frac{8}{14} = 0.032$$

Slightly Different Warm-Up Exercise

Question: How would we classify  ?



Need to compute $p(- \mid \mathbf{x} = [\text{green}, \text{circle}]) \propto$

$$\hat{p}(\text{green} \mid -) \cdot \hat{p}(\text{circle} \mid -) \cdot \hat{p}(-) =$$

$$\frac{2+1}{5+3} \cdot \frac{0+1}{5+2} \cdot \frac{5+1}{12+2}$$

$$= \frac{3}{8} \cdot \frac{1}{7} \cdot \frac{6}{14} = .023$$

Slightly Different Warm-Up Exercise

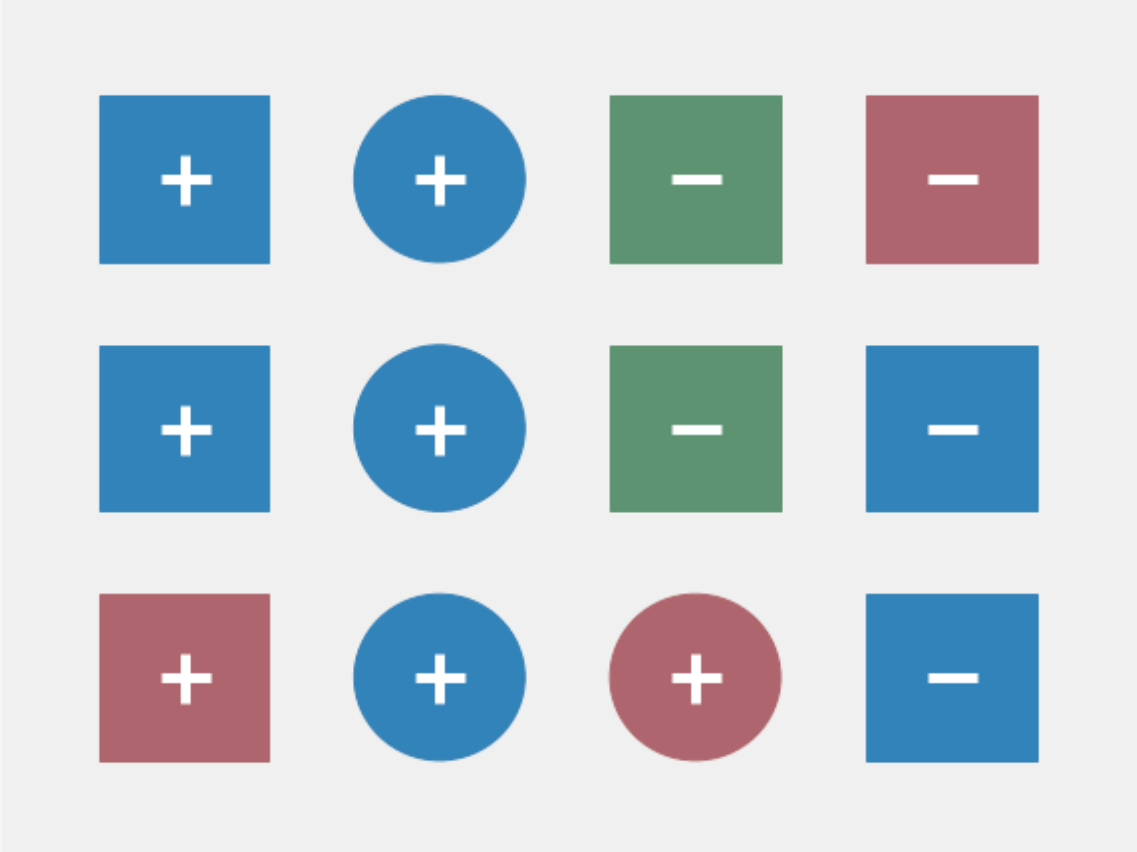
Question: How would we classify  ?

So we have:

$$p(+ \mid \mathbf{x} = [\text{green, circle}]) \propto 0.032$$

$$p(- \mid \mathbf{x} = [\text{green, circle}]) \propto 0.023$$

So we predict positive



Slightly Different Warm-Up Exercise

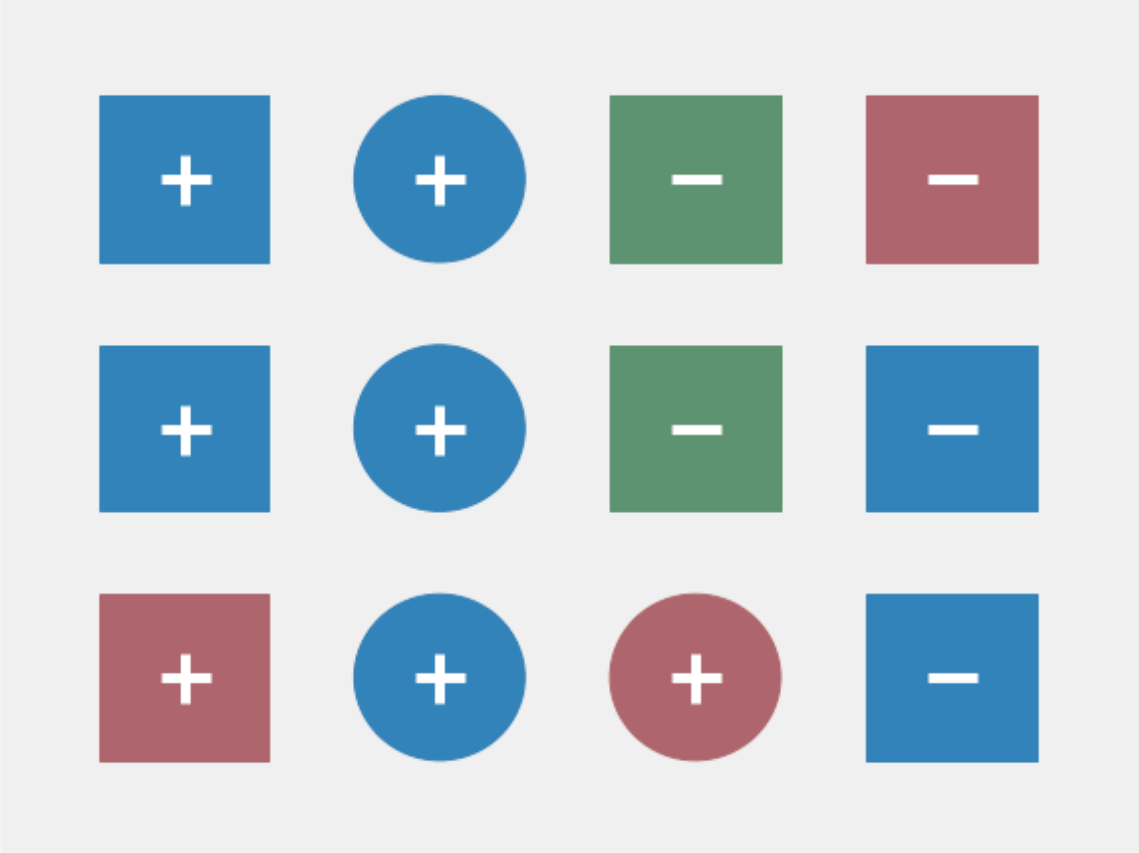
Question: How would we classify  ?

So we have:

$$p(+ \mid \mathbf{x} = [\text{green, circle}]) \propto 0.032$$

$$p(- \mid \mathbf{x} = [\text{green, circle}]) \propto 0.023$$

So we predict positive .



Slightly Different Warm-Up Exercise

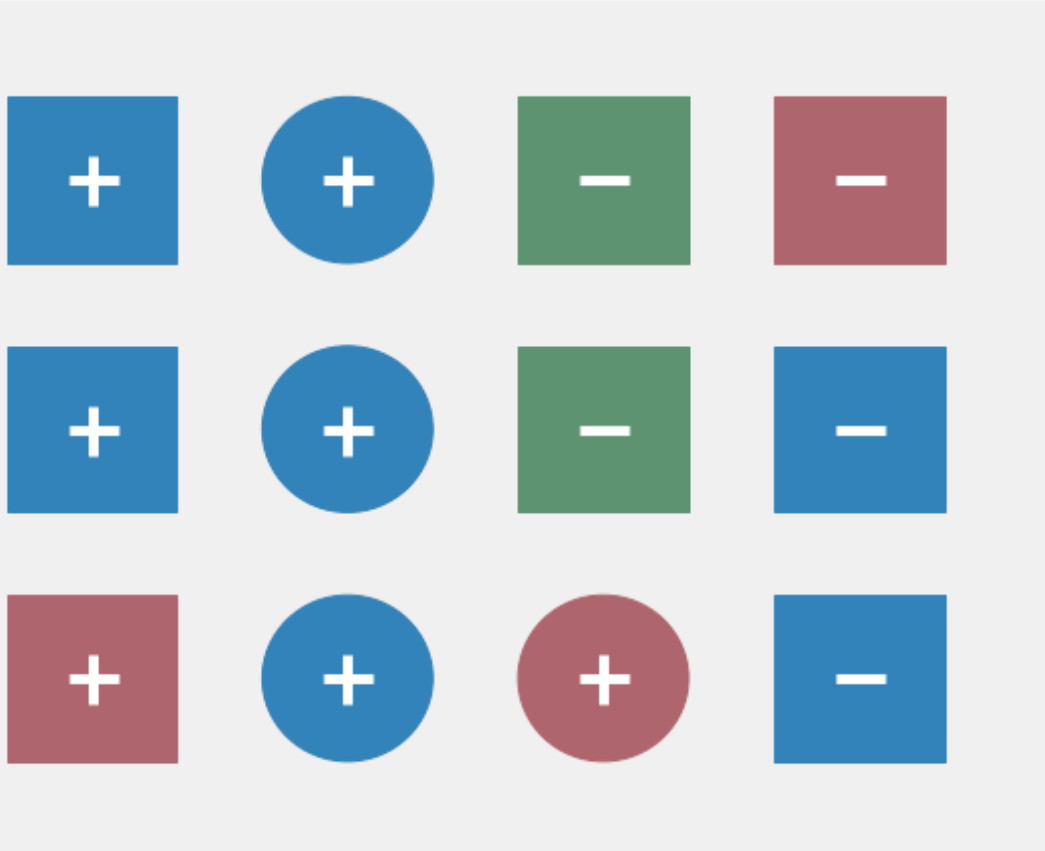
Question: How would we classify  ?

Uhhh…

$$P(\text{wiggly} \mid +) = \frac{0+1}{7+2+1} = \frac{1}{10}$$

$$P(\square \mid +) = \frac{3+1}{7+2+1} = \frac{4}{10}$$

$$P(\text{o} \mid +) = \frac{4+1}{7+2+1} = \frac{5}{10}$$



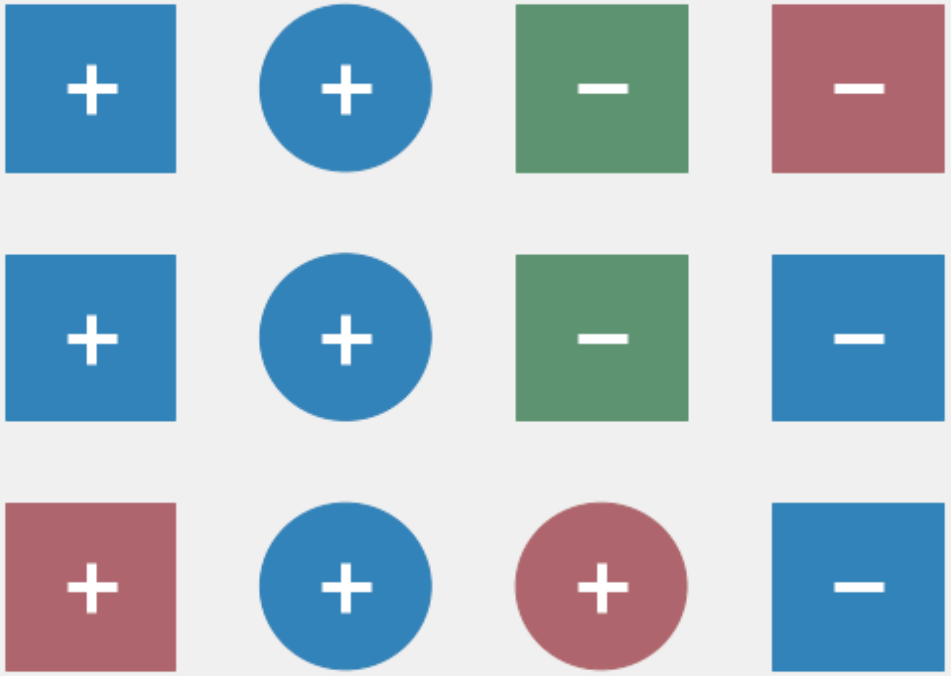
Slightly Different Warm-Up Exercise

Question: How would we classify  ?

Orange doesn't appear in training set

Add **UNKCOLOR** feature value

While we're at it, add **UNKSHAPE** feature too



Slightly Different Warm-Up Exercise

Question: How would we classify  ?

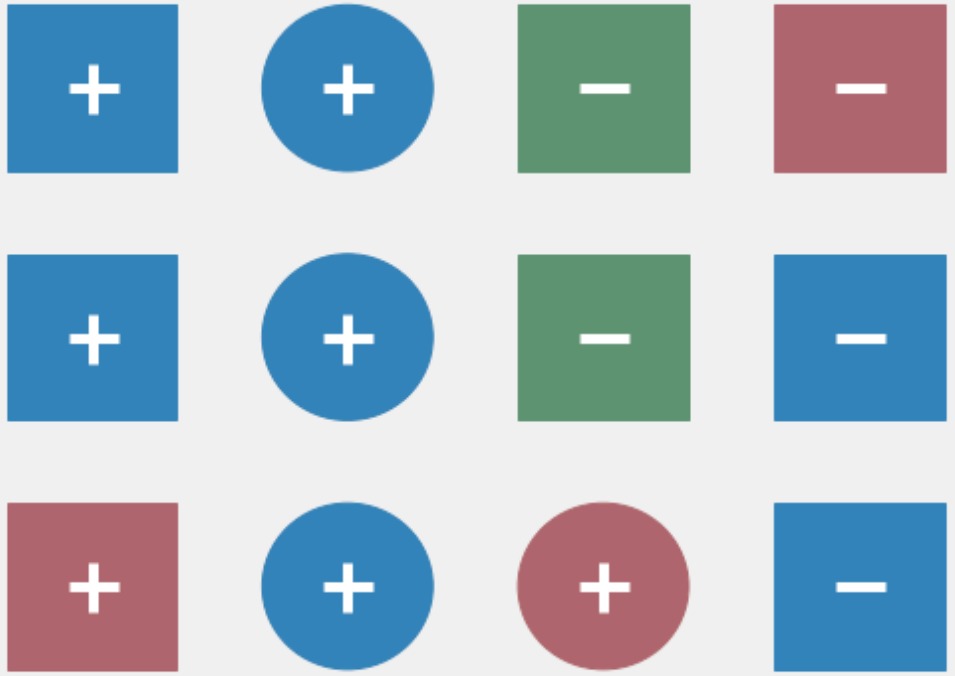
Need $p(+ \mid \mathbf{x} = [\text{orange, square}]) \propto$

$$\hat{p}(\text{orange} \mid +) \cdot \hat{p}(\text{square} \mid +) \cdot \hat{p}(+)$$

$$\frac{0+1}{7+4} \cdot \frac{3+1}{7+3} \cdot \frac{7+1}{12+2}$$

$$= 0.021$$

UNKCOLOR, UNKSHAPE

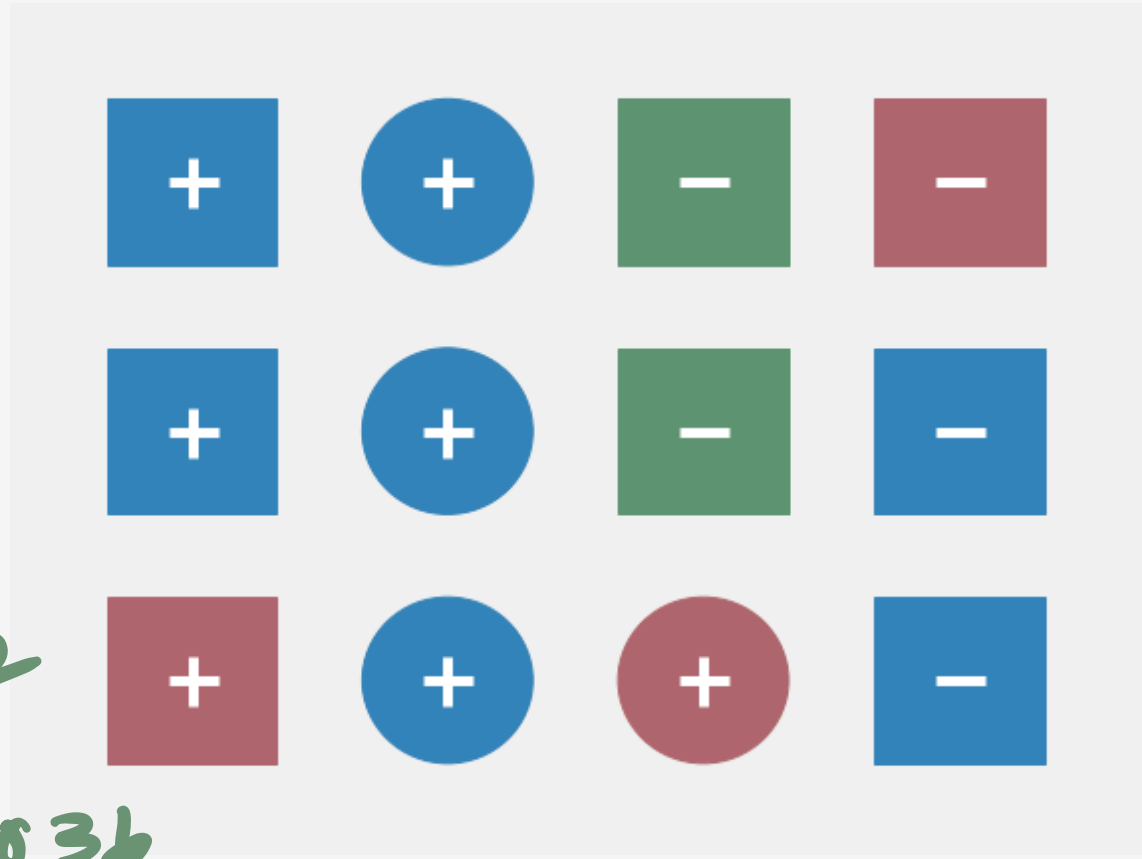


$V_S = \{$ CIR, SQR, UNKSHAPE $\}$

$V_C = \{$ RED, GREEN, BLUE, UNKCOLOR

Slightly Different Warm-Up Exercise

Question: How would we classify  ?



Need $p(- \mid \mathbf{x} = [\text{orange, square}]) \propto$

$$\hat{p}(\text{orange} \mid -) \cdot \hat{p}(\text{square} \mid -) \cdot \hat{p}(-)$$

$$\frac{0+1}{5+4} \cdot \frac{5+1}{5+3} \cdot \frac{5+1}{12+2}$$

$$= \frac{1}{9} \cdot \frac{6}{8} \cdot \frac{6}{14} = 0.036$$

Slightly Different Warm-Up Exercise

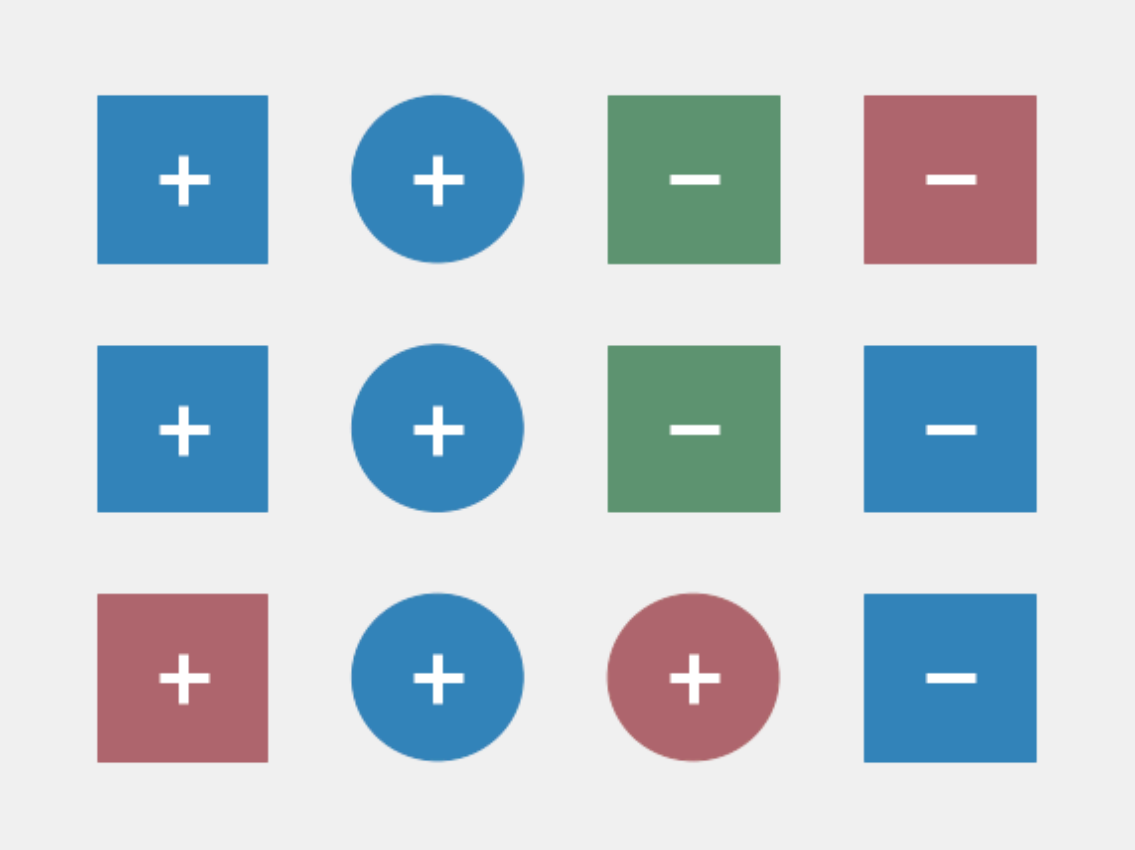
Question: How would we classify  ?

So we have:

$$p(+ \mid \mathbf{x} = [\text{orange, square}]) \propto 0.021$$

$$p(- \mid \mathbf{x} = [\text{orange, square}]) \propto 0.036$$

So we predict NEGATIVE



Slightly Different Warm-Up Exercise

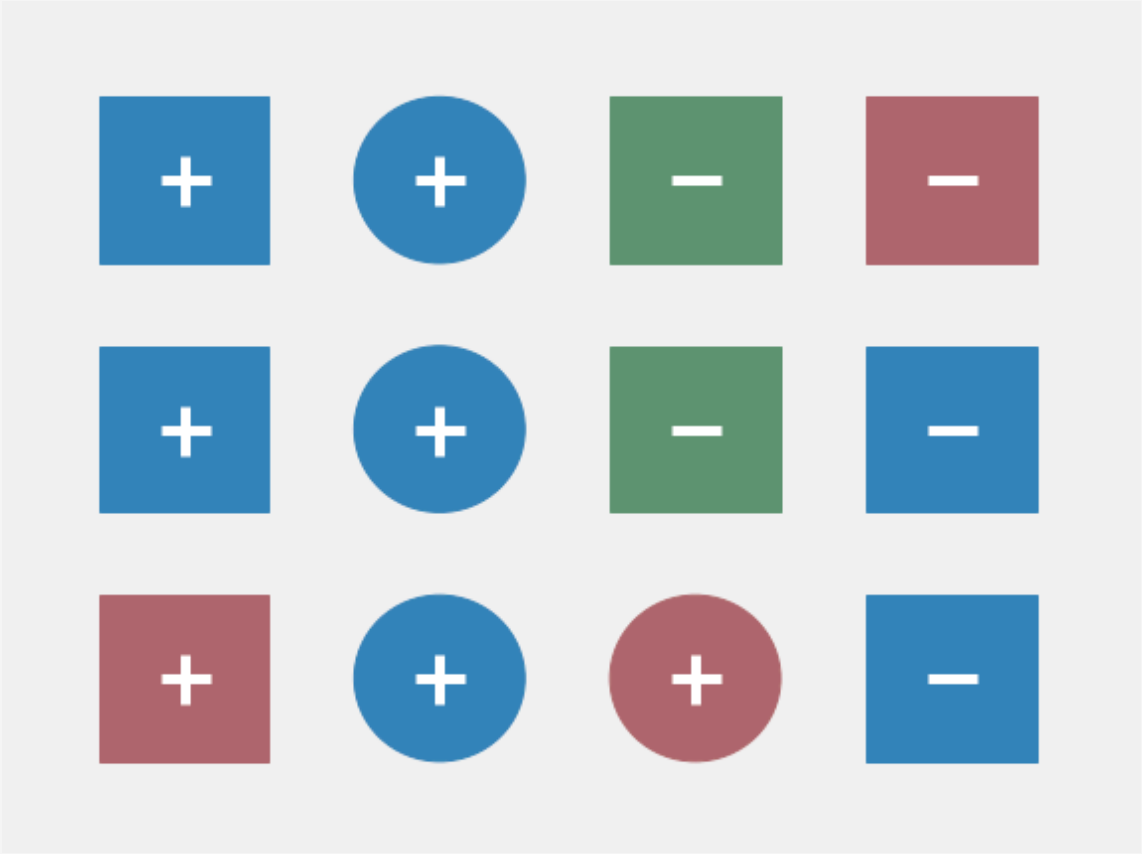
Question: How would we classify  ?

So we have:

$$p(+ \mid \mathbf{x} = [\text{orange, square}]) \propto 0.021$$

$$p(- \mid \mathbf{x} = [\text{orange, square}]) \propto 0.036$$

So we predict negative .



Naïve Bayes

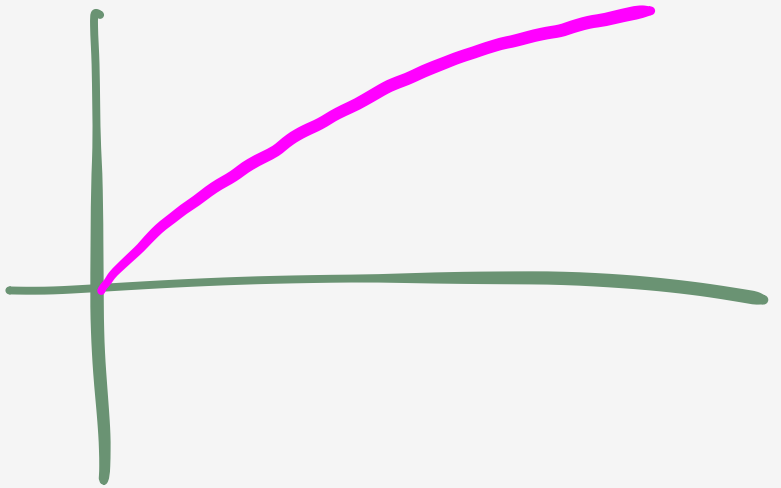
- The Naïve Bayes classifier is a probabilistic classifier
- We compute the posterior score of message $\mathbf{x}$ belonging to class c as

$$p(y = c \mid \mathbf{x}) \propto \hat{p}(c) \prod_k \hat{p}(x_k \mid c)$$

- Predicted class is the one with the highest posterior score

$$\hat{y} = \arg\max_c p(y = c \mid \mathbf{x}) = \arg\max_c \hat{p}(c) \prod_k \hat{p}(x_k \mid c)$$

Numerical Hiccup



For long messages, have to multiply a lot of probabilities together

$$\hat{y} = \arg\max_c \hat{p}(c) \prod_k \hat{p}(x_k \mid c)$$

Probabilities are numbers less than 1. If you multiply enough of them, you could get **underflow**

Fix: Compute and store the **log** of the term-class score estimates

Recall: The log of a product is the sum of the logs: $\log(ab) = \log(a) + \log(b)$

Predicted class becomes: $\hat{y} = \arg\max_c \log \hat{p}(c) + \sum_k \log \hat{p}(x_k \mid c)$

Training and Prediction Complexity

Question: How much does Naïve Bayes cost to train?

Training and Prediction Complexity

Question: How much does Naïve Bayes cost to train?

Answer: It's linear in the number of features and number of training examples!

- Single pass over training data to do frequency counts
- Another pass over recorded frequencies to normalize and take logs

Question: How much does Naïve Bayes cost to make a prediction?

Training and Prediction Complexity

Question: How much does Naïve Bayes cost to train?

Answer: It's linear in the number of features and number of training examples!

- Single pass over training data to do frequency counts
- Another pass over recorded frequencies to normalize and take logs

Question: How much does Naïve Bayes cost to make a prediction?

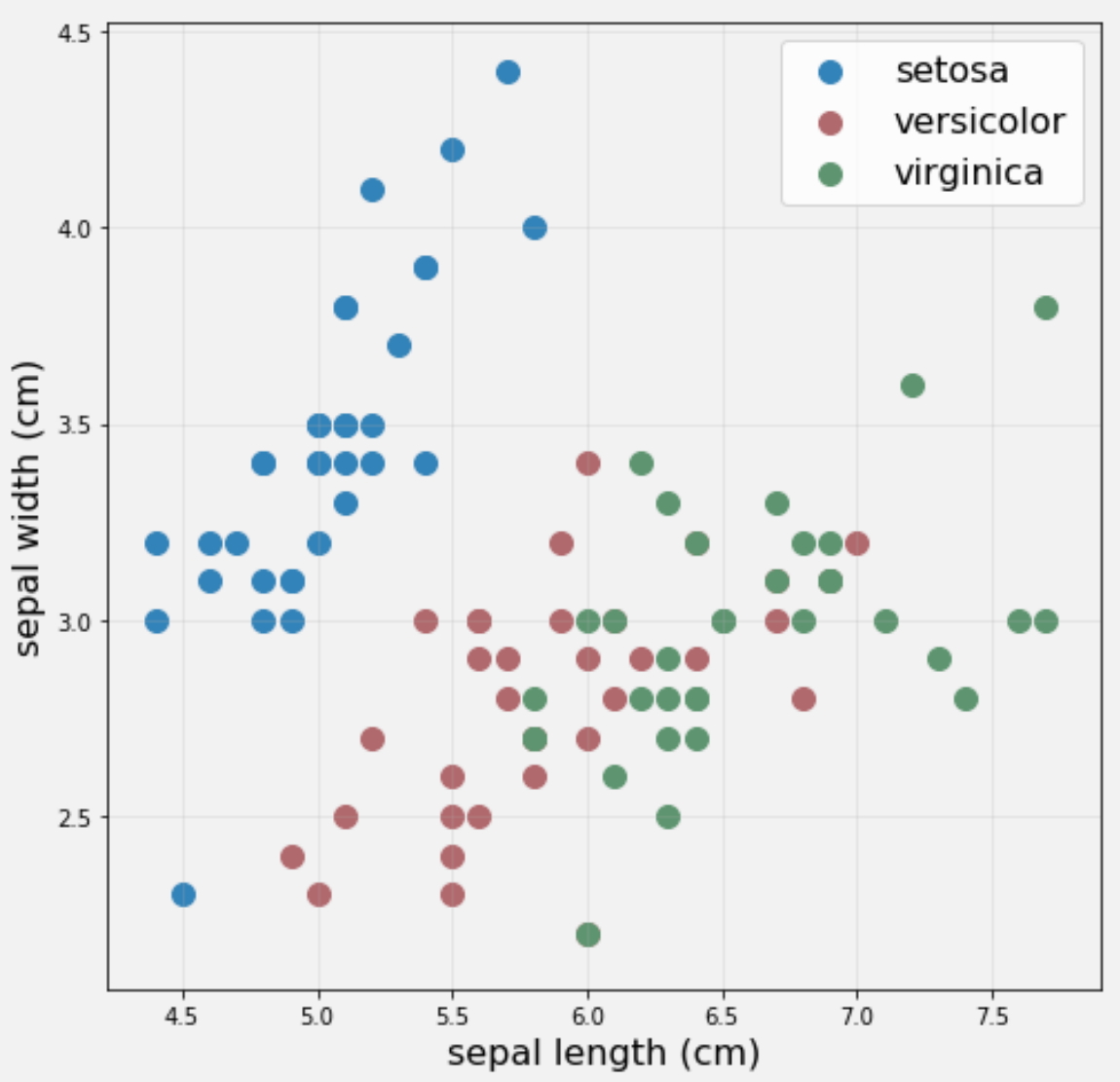
Answer: It's linear in the number of features present in the query point!

Naïve Bayes with Continuous Features

Example: Predict Iris species from sepal length and sepal width.

Features are now continuous values instead of frequencies…

What do we do?!



Naïve Bayes with Continuous Features

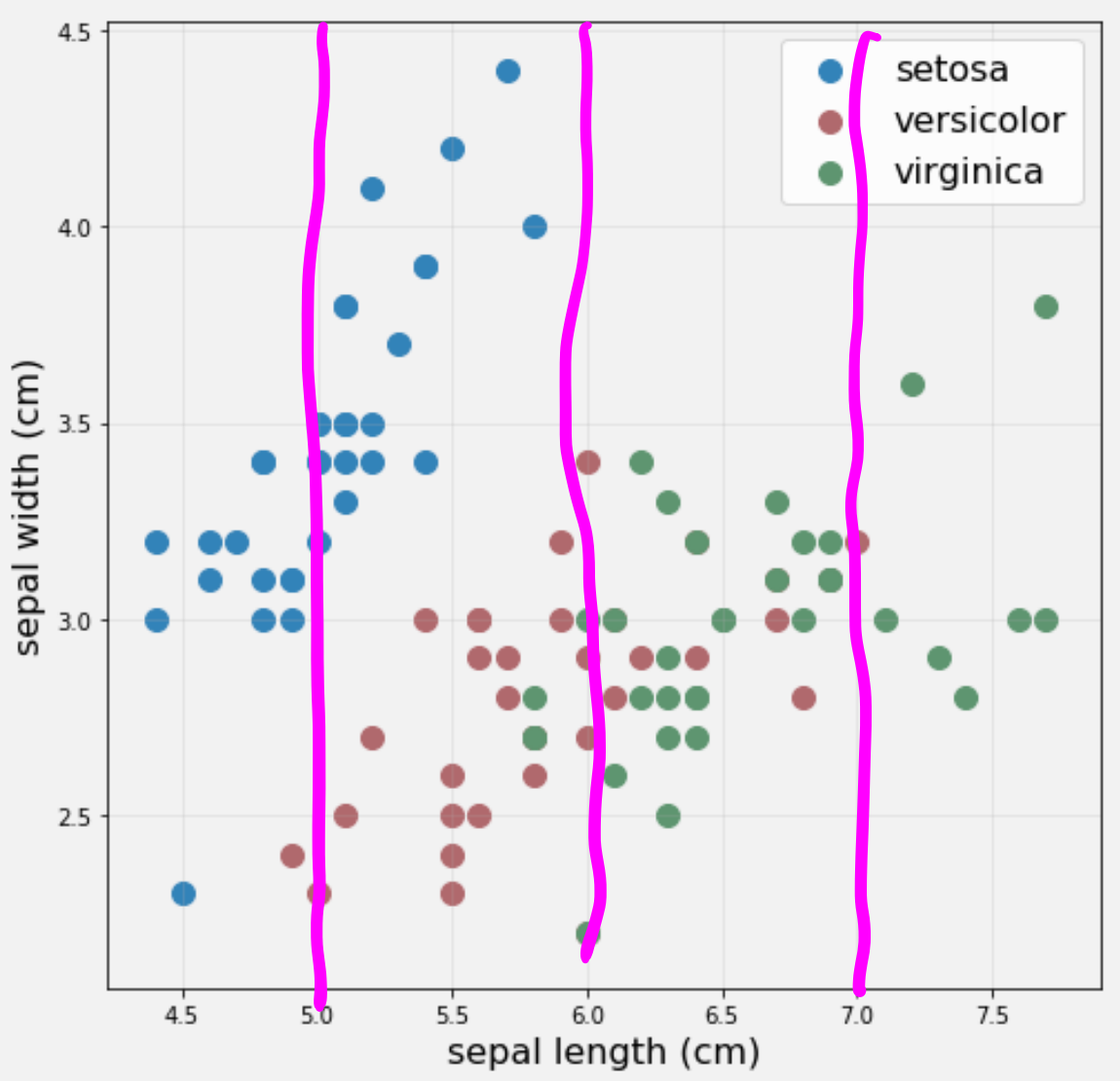
Example: Predict Iris species from sepal length and sepal width.

Features are now continuous values instead of frequencies...

Idea: Discretize the Continuous Features

$$x_k = \begin{cases} 0 & \text{if } x_k < Q_1 \\ 1 & \text{if } x_k \in [Q_1, Q_2] \\ 2 & \text{if } x_k \in [Q_2, Q_3] \\ 3 & \text{if } x_k > Q_3 \end{cases}$$

Now just pretend features are frequency counts



Gaussian Naïve Bayes

Idea: Assume features follow a normal dist

- Priors don't change.

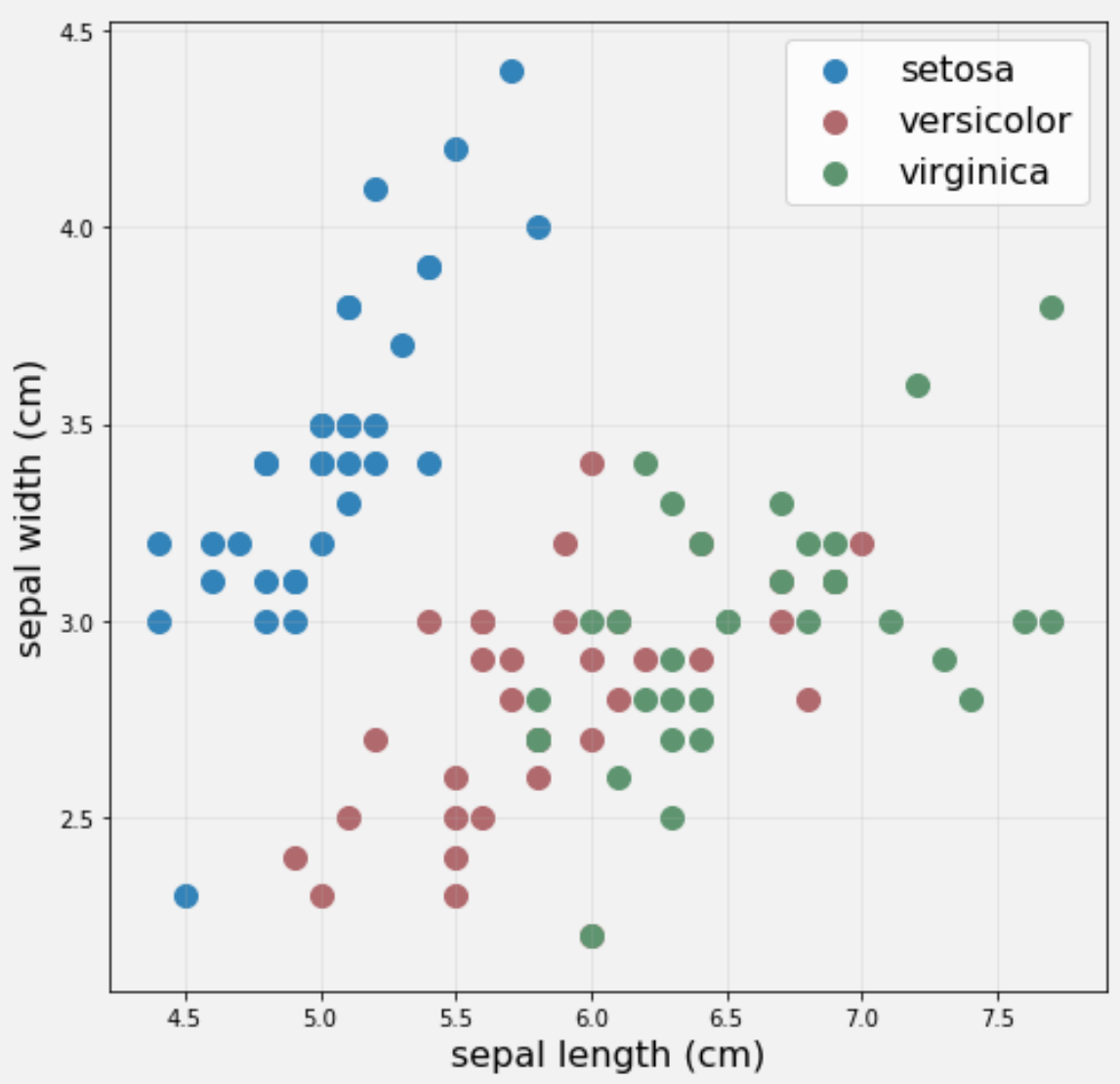
$$\hat{p}(y = c) = \frac{1}{n} \sum_{i=1}^{n} I(y_i = c)$$

- Likelihoods are computed from Gaussian pdf

S. LENGTH

$$p(X_k = x \mid Y = c) =$$

$$\frac{1}{\sqrt{2\pi\sigma_{kc}}} \exp\left[-\frac{1}{2} \left(\frac{x - \mu_{kc}}{\sigma_{kc}} \right)^2 \right]$$



Gaussian Naïve Bayes

- Likelihoods are computed from Gaussian pdf

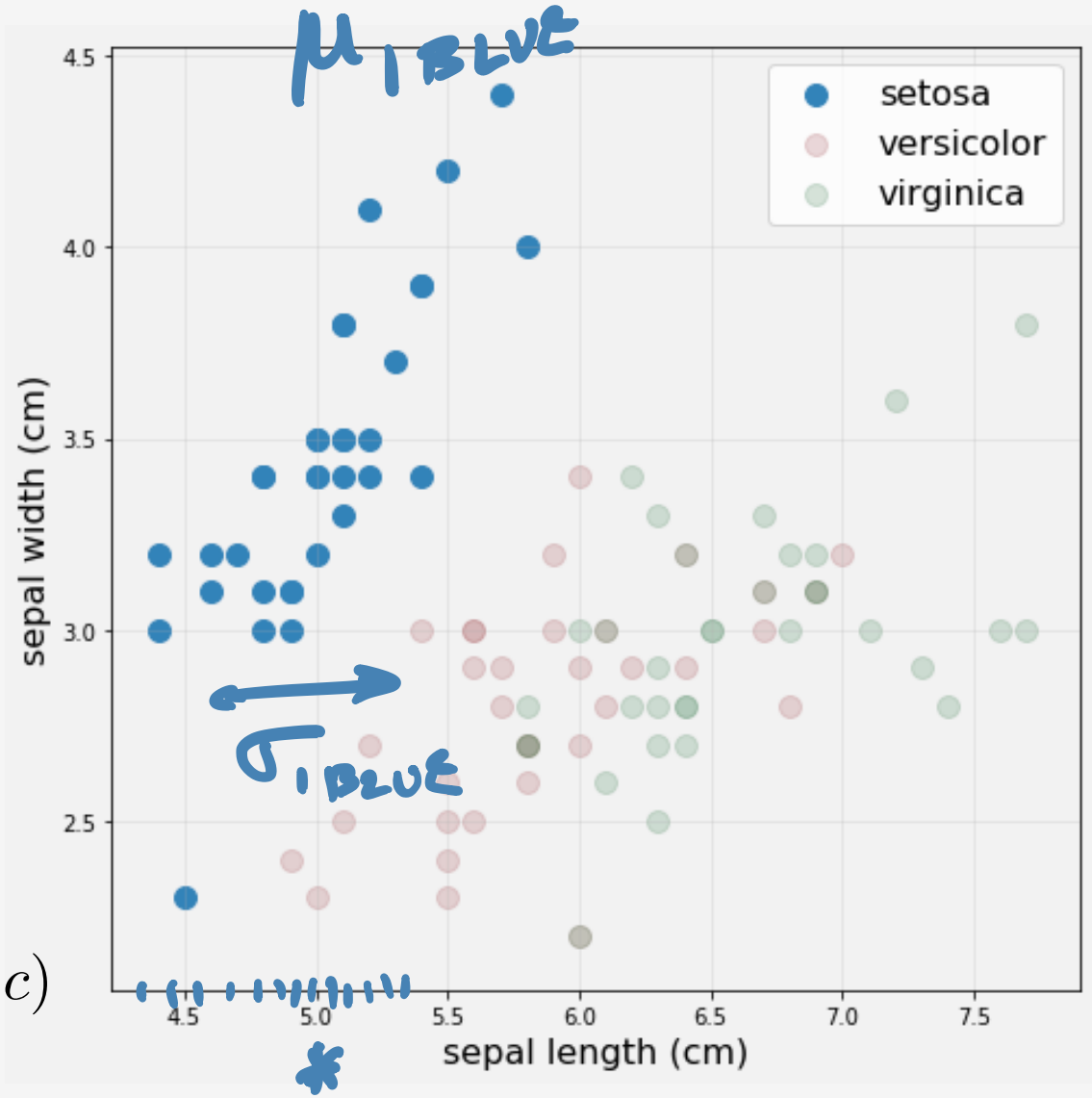
$$p(X_k = x \mid Y = c) =$$

$$\frac{1}{\sqrt{2\pi\sigma_{kc}}} \exp\left[-\frac{1}{2}\left(\frac{x - \mu_{kc}}{\sigma_{kc}}\right)^2\right]$$

- Get MLE estimates of params from training set

$$\mu_{kc} = \frac{1}{\sum_{i=1}^{n} I(y_i = c)} \sum_{i=1}^{n} x_{ik} I(y_i = c)$$

$$\sigma_{kc}^2 = \frac{1}{\sum_{i=1}^{n} I(y_i = c)} \sum_{i=1}^{n} (x_{ik} - \mu_{kc})^2 I(y_i = c)$$

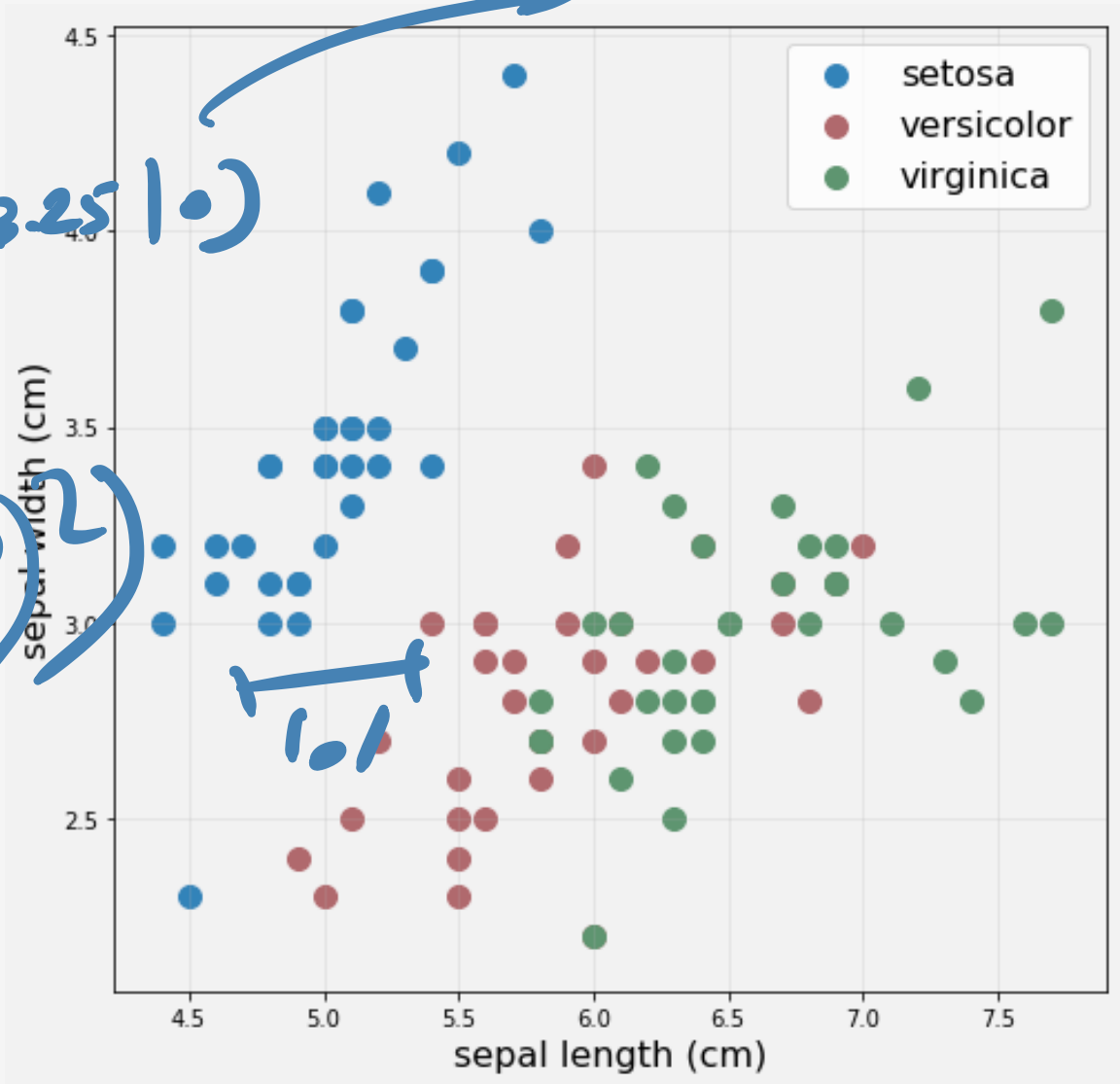


Gaussian Naïve Bayes

$$P(\bullet \mid SL = 6.5, SW = 3.25)$$

$$\propto \hat{P}(\bullet) \times P(SL = 6.5 \mid \bullet) \times P(SW = 3.25 \mid \bullet)$$

$$\frac{1}{\sqrt{2\pi}\,\sigma_{i,blue}} \exp\left[-\frac{1}{2}\left(\frac{6.5 - \mu_{i,blue}}{\sigma_{i,blue}}\right)^2\right]$$



$$p(\Delta \mid +) =$$