Follow Torch the triceratops, our new class mascot on instagram: @torchthetriceratops

1 Definitions Woohoo!

Naive Bayes: Naive Bayes is a probabilistic algorithm that uses Bayes' theorem to classify data.

1. **Naive Bayes Assumption**: features of a data point are conditionally independent given the class label. That is all X_i are conditionally independent given Y or

$$P(X_1, X_2, \ldots, X_i \mid Y) = \prod_{i=1}^{I} P(X_i \mid Y)$$

2. **Bayes Theorem**:

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

3. **Naive Bayes Algorithm**:
   (a) Calculate the prior probability of each class label.
   (b) For each feature x_i of each class label, calculate the conditional probability $P(x_i \mid y)$.
   (c) Given a new data point x, calculate the posterior probability of each class label using Bayes' theorem and the conditional probabilities from step b. The class label with the highest posterior probability is the predicted class label.

Generative Models: class of models that can generate new data points that are similar to the training data. That is, generative models model $P(Y \mid \theta_{\text{class}})$ and $P(X \mid Y, \theta_{\text{class conditional}})$

The parameters θ_{class} and $\theta_{\text{class conditional}}$ are learned from the data. Bayes rule can then be used to compute $P(Y \mid X, \theta)$

$$P(Y \mid X, \theta) \propto P(X \mid Y, \theta_{\text{class conditional}})P(Y \mid \theta_{\text{class}})$$

Properties:

- More model assumptions involved
- Rich model allows for generating new data points that are similar to the training data.
- Can be used for unsupervised learning, where the class labels are not known.

Discriminative Models: Discriminative models model $P(Y \mid X, \theta)$ directly. The parameters, θ, are also learned from the data.

Properties:

- Less model assumptions needed
- Requires more labeled data

1.1 Regularization and MAP

	Regularization Penalty	**Prior**
Ridge Regression	$\|\|\mathbf{w}\|\|_2^2$	$w_j \sim \mathcal{N}(0, \tau^2)$
Lasso	$\|\|\mathbf{w}\|\|_1$	$w_j \sim \text{Laplace}(0, b)$

2 Comparing models

	MLE	**MAP**
Discriminative	• Linear Regression • Logistic Regression • Logistic Regression with polynomial features	• Linear regression with L2 regularization • Logistic Regression with Laplace Prior
Generative	• Naive Bayes	• Naive Bayes with Laplace smoothing*

Consider the list of models we've learned so far and place them in the correct boxes above:

- Linear regression
- Logistic regression
- Linear regression with L2 regularization
- Logistic regression with Laplace prior
- Logistic regression with polynomial features
- Naive Bayes

3 Naive Bayes

Consider the following training samples:

SPAM	Email Body
1	Money is free now
0	Pat teach 315
0	Pat free to teach
1	Sir money to teach
1	Pat free money now
0	Teach 315 now
0	Pat to teach 315

The vocabulary consists of the following words: {315, free, is, money, now, Pat, Sir, teach, to, tomorrow}. Compute the following probabilities:

1. Fill in the tables below:

$P(Y=1)$	$P(Y=0)$

$P(Y=1) = 3/7$

$P(Y=0) = 4/7$

j	$P(X_j = 1 \mid Y = 1)$	$P(X_j = 1 \mid Y = 0)$
315	0/3	3/4
free	2/3	1/4
is	1/3	0/4
money	3/3	0/4
now	2/3	1/4
Pat	1/3	3/4
Sir	1/3	0/4
teach	1/3	4/4
to	1/3	2/4
tomorrow	0/3	0/4

2. Consider the following email body: X = Pat teach now. Fill in the table below:

j	x	$P(X_j = x \mid Y = 1)$	$P(X_j = x \mid Y = 0)$
315	0	3/3	1/4
free	0	1/3	3/4
is	0	2/3	4/4
money	0	0/3	4/4
now	1	2/3	1/4
Pat	1	1/3	3/4
Sir	0	2/3	4/4
teach	1	1/3	4/4
to	0	2/3	2/4
tomorrow	0	3/3	4/4

3. Reminder that with naive Bayes, $P(Y, X_1, \ldots, X_M) = \prod P(X \mid Y)P(Y)$

$P(Y = 1, X_1, ..., X_M)$	$P(Y = 0, X_1, ..., X_M)$

$P(Y = 1, X_1, ..., X_m) = \prod P(X \mid Y = 1)P(Y = 1)$
3/3 * 1/3 * 2/3 * 0/3 * 2/3 * 1/3 * 2/3 * 1/3 * 2/3 * 3/3 * 3/7
=0

$P(Y = 0, X_1, ..., X_M)$
1/4 * 3/4 * 4/4 * 4/4 * 1/4 * 3/4 * 4/4 * 4/4 * 2/4 * 4/4 * 4/7
=0.01004

4.

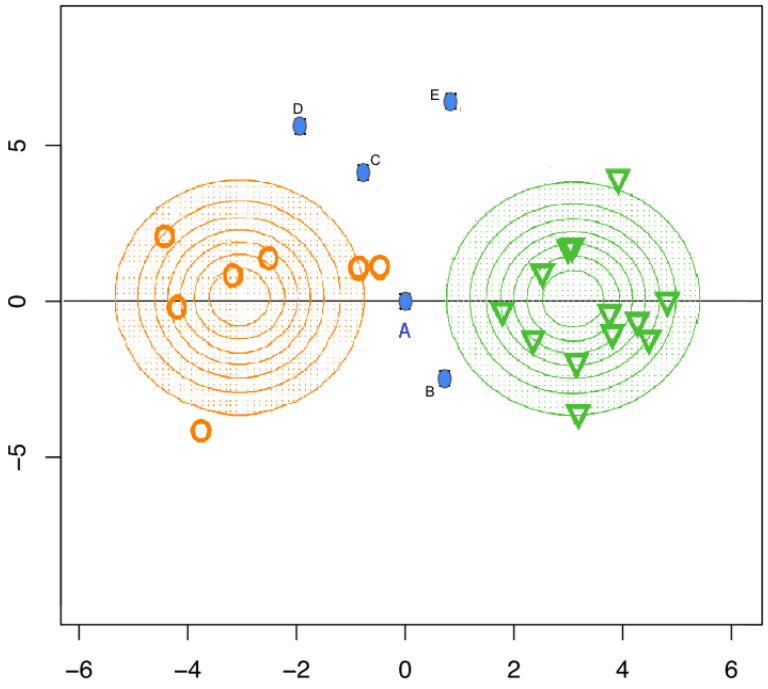
$P(Y = 1 \mid X_1, ..., X_M)$	$P(Y = 0 \mid X_1, ..., X_M)$
$0 \ / \ (0.01004 + 0) = 0$	$0.01004 \ / \ (0.01004 + 0) = 1$

4 Gaussian Discriminant Analysis (A generative method)

Gaussian discriminant analysis is used when the input features are continuous and $p(\mathbf{x} \mid y)$ is modeled as a multivariate Gaussian distribution.
Note: Since we're dealing with $p(\mathbf{x} \mid y)p(y)$, it is a generative model, despite its name!

a. Consider two Gaussian distributions as formulated and visualized below:

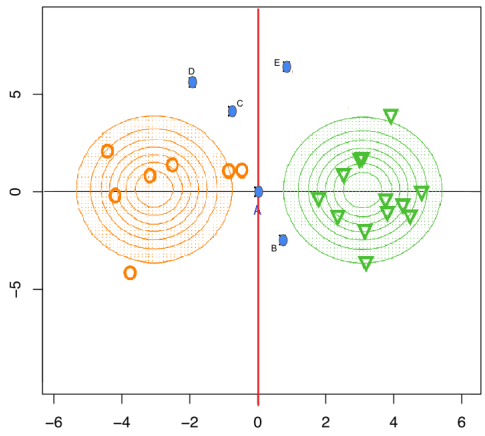
$\mathbf{x} \mid y = 0 \sim \mathcal{N}(\mu_{y=0}, \mathbf{I})$
$\mathbf{x} \mid y = 1 \sim \mathcal{N}(\mu_{y=1}, \mathbf{I})$



i. X are some observed data points. You are given that point A lies on the midpoint between the two distribution centers. Label each data point with its likely class.
(hint: Both distributions have the same covariance!)

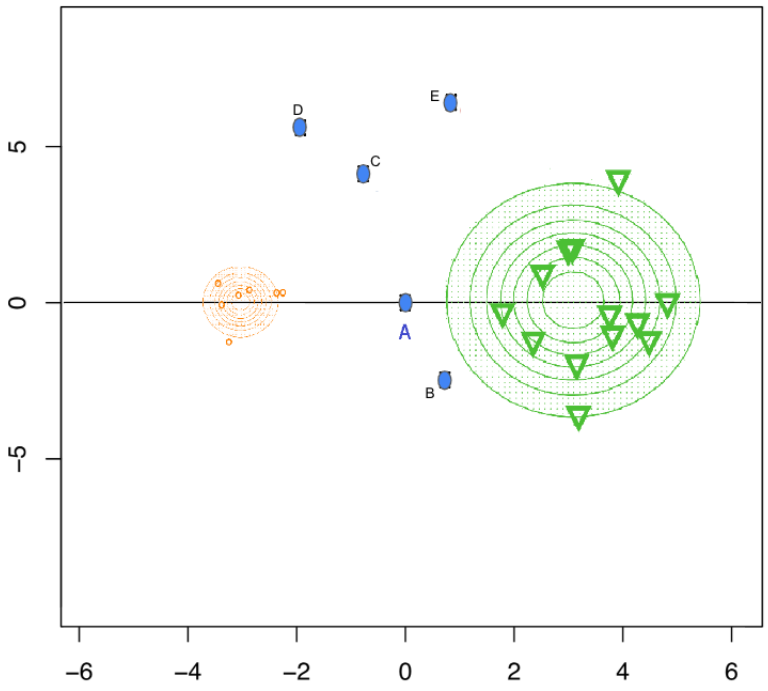
B - Green, C - Orange, D - Orange, E - Green

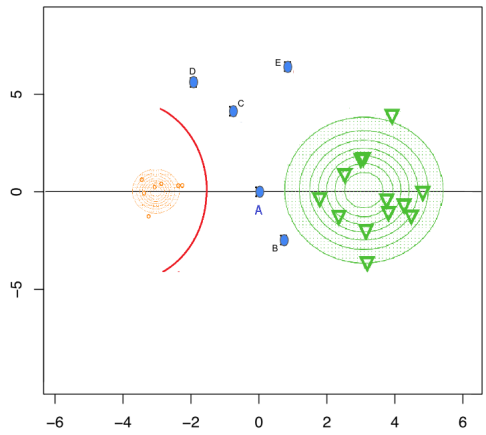
ii. If we were to draw a boundary separating the two distributions, where would the boundary be and what would it look like?



The boundary will be linear.

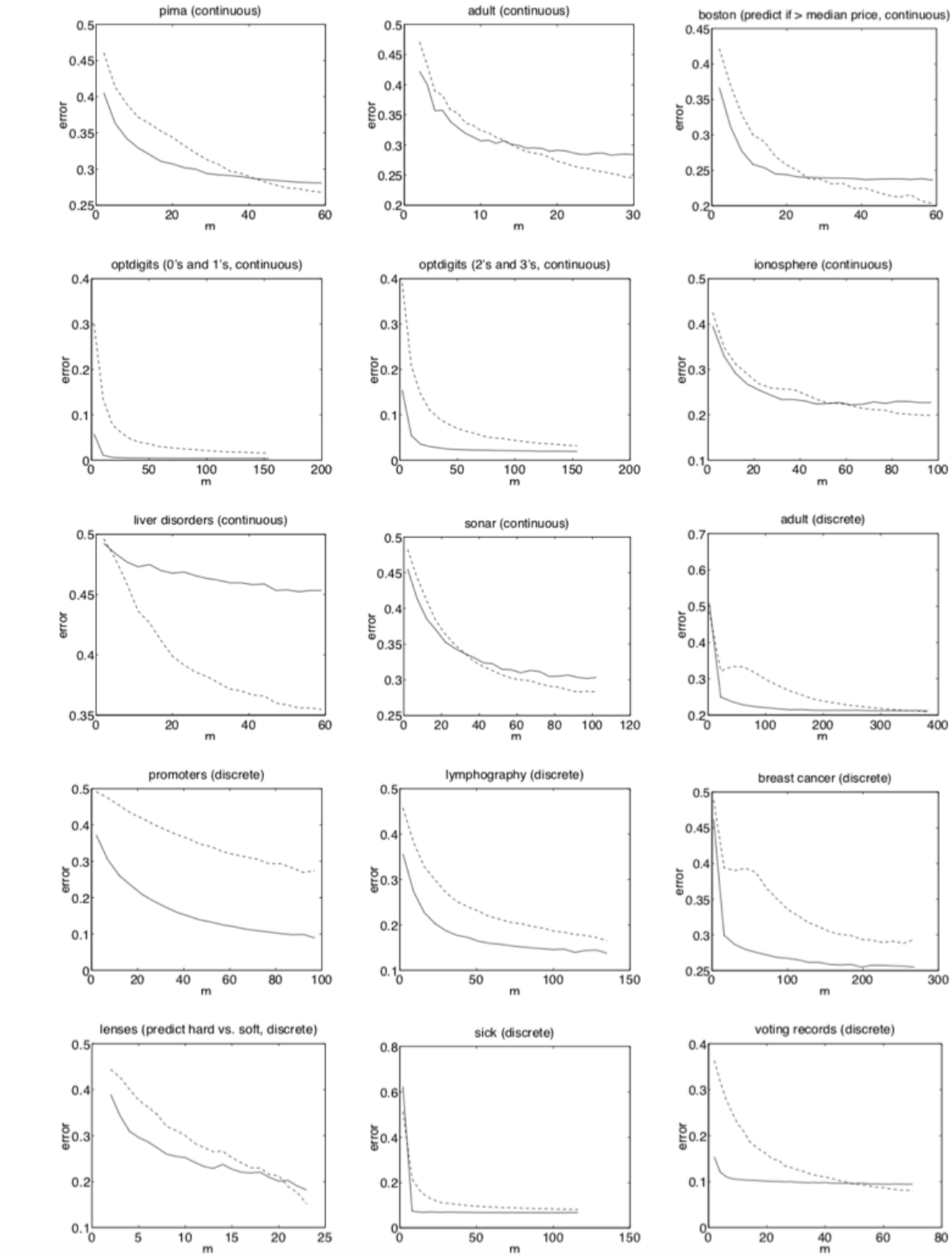
b. Now supposed a different distrbution, whose covariance matrix is $\frac{1}{5}$I. Re-label the data points again. Point A lies on the midpoint between the two distribution centers. How does the boundary change?





The boundary will now be the zero-crossing of a quadratic function or an ellipse about the orange distribution (only a segment of it is drawn above).

5 Performance of Generative vs Discriminative models



These are 15 experiments run by Andrew Ng and Michael Jordan at U.C. Berkeley using 15 standard UCI data sets, in order to compare the performance of Generative vs Discriminative models in real applications. (FYI, this is the link to the original paper: https://ai.stanford.edu/~ang/papers/nips01-discriminativegenerative.pdf) In each of the plots, the X-axis is number of samples, and Y-axis is average error across 1,000 random splits of training/validation set. Look at the graphs and answer the following:

a. The two models considered are Naive Bayes for Generative model and Logistic Regression for Discriminative model, where the dashed line represents Logistic Regression and the solid line represents Naive Bayes. According to these results, which model has the better asymptotic performance?

> In most cases, Logistic Regression has the better asymptotic performance.

b. What performance advantages does each type of model possess?

> Generative: better performance when data is scarce. Additional ability to simulate new samples, because it models the joint distribution of $\mathbf{x}$ and y.
>
> Discriminative: better asymptotic performance. Better performance on high-dimensional feature space with dependent features.