1 Let's Generative Some Definitions!

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 $P(X \mid Y, \theta_{\text{class conditional}})$. In addition to generating new data, generative models can also be used for prediction.

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

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

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 Generative + MAP

We can also combine generative models with MAP by adding a prior on the parameters:

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

1.2 Naive Bayes Recap

Naive Bayes is a probabilistic model that simplifies a generative model by using the naive Bayes assumption, which states that all input features of a data point are conditionally independent from each other given the output value. That is, for all $i \neq j$, X_i and X_j are conditionally independent given Y or:

$$P(X_1, X_2, ..., X_j \mid Y) = \prod_{j=1}^{M} P(X_j \mid Y)$$

Naive Bayes Optimization:

1. Similar to MLE, estimate the parameters for the class prior distribution, $P(Y \mid \theta_{class})$.
2. **For each class label** y, use the data points with label y only to estimate the parameters (similar to MLE) for $P(X_j \mid Y = y, \theta_{\text{class conditional},y,j})$ **for each feature** X_j**, independently**.

Naive Bayes Inference (Prediction): Given a new set of input features $x_1, \cdot, x_M$, class label with the highest posterior probability, $P(Y = y \mid x_1, \cdot, x_M)$ by taking advantage of Bays theorem:

$$\underset{y}{\text{argmax}}\, P(Y = y) \prod_{j=1}^{M} P(X_j \mid Y = y)$$

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

2.1 Reminder: Regularization and MAP

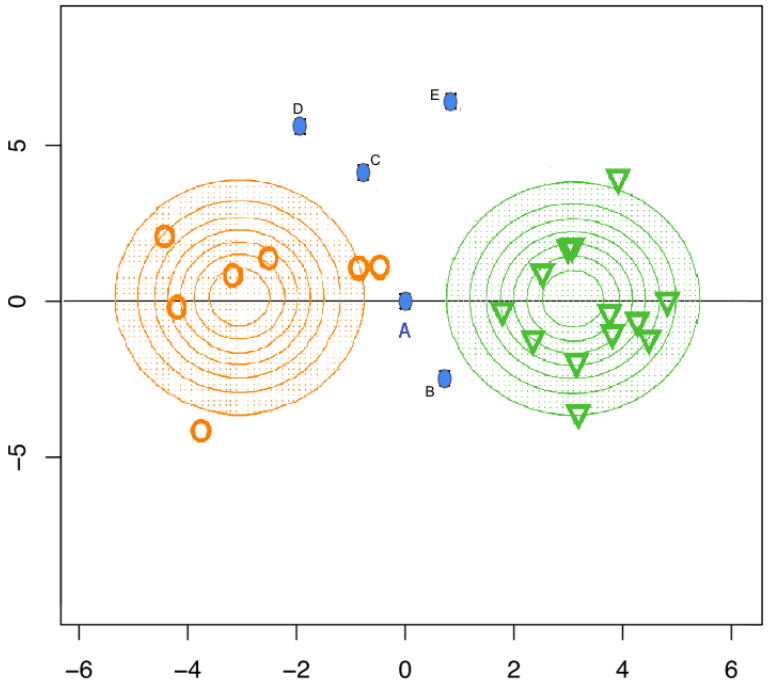
	Regularization Penalty	**Prior**
Ridge Regression	$\|\|\mathbf{w}\|\|_2^2$	$w_j \sim \mathcal{N}(0, \tau^2)$
Lasso	$\|\|\mathbf{w}\|\|_1$	$w_j \sim \text{Laplace}(0, b)$

3 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:

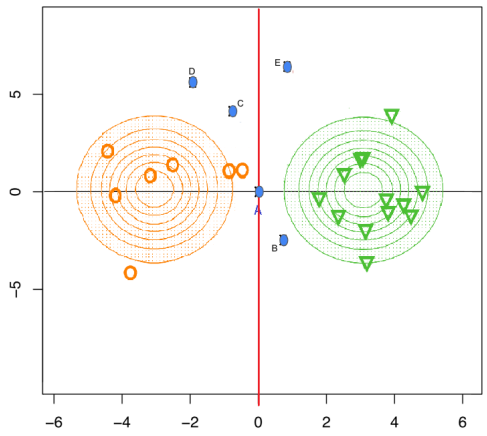
$\mathbf{x} \mid y = 0 \sim \mathcal{N}(\mu_{y=0}, \mathbf{I})$
$\mathbf{x} \mid y = 1 \sim \mathcal{N}(\mu_{y=1}, \mathbf{I})$



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!)

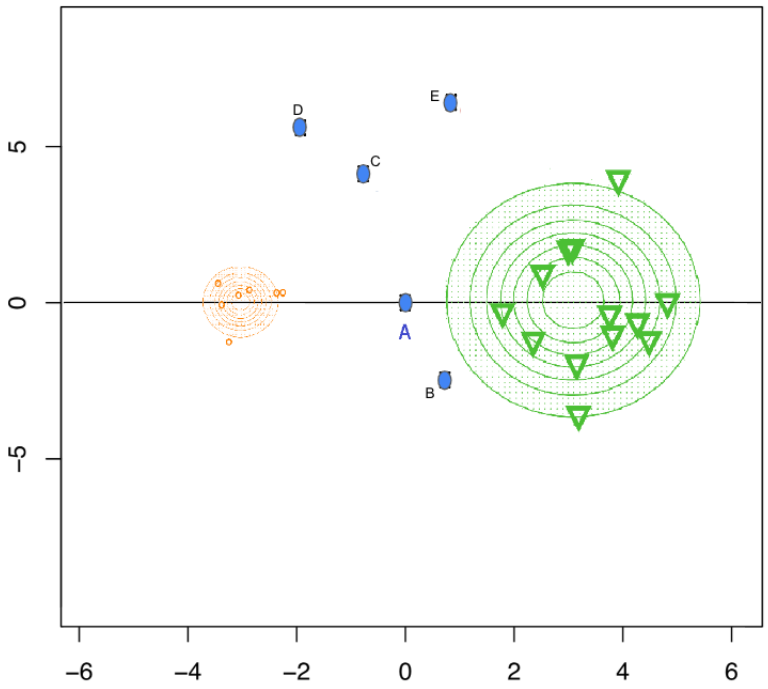
B - Green, C - Orange, D - Orange, E - Green

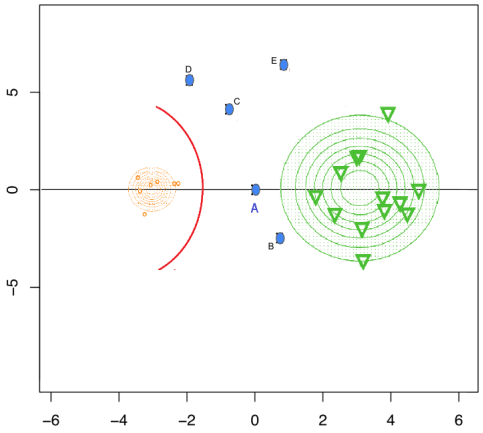
ii. If we were to draw a boundary separating the two distributions, where would the boundary be and what would it look like?



The boundary will be linear.

b. Now supposed a different distribution, 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).

4 Naive Bayes with Laplace Smoothing

4.1 Naive Bayes

Naive Bayes is a probabilistic algorithm that uses Bayes' theorem to classify data. Reminder that Bayes Theorem states

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

In addition to Bayes Theorem, Naive Bayes makes use of something called the Naive Bayes assumption, which states that features of a data point are conditionally independent given the class label. That is, all X_i are conditionally independent given Y or

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

The **Naive Bayes Algorithm** works as follows:

1. Calculate the prior probability of each class label, $P(Y = y)$.
2. For each feature X_i of each class label, calculate the conditional probability $P(X_i \mid Y = y)$.
3. 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.

   In other words, assuming there are M features, the classifier can be written as

$$\underset{y}{\operatorname{argmax}}\, P(Y = y) \prod_{i=1}^{M} P(X_i \mid Y = y)$$

4.2 Laplace Smoothing

It's very convenient to use the Naive Bayes assumption that the featues are independent given the class–it allows us to simplify the calculations a lot. But, whenever a word/token doesn't occur for a given class, the probability for that token given that class becomes 0, hence the probability of that class also becomes 0. This is a massive drawback. We cannot extrapolate our model to an example that contains a word that we have not seen before in our training data. If we tried, the probability for each class would just be 0.

Laplace smoothing is a technique used to address this issue by adding a small pseudo-count (denoted by α) to each observation. This ensures that no probability estimation becomes zero.

To be more precise,

$$P(x_i = k|Y = y) = \frac{(\text{Number of samples that belong to class y and in which } x_i \text{ takes on class k}) + \alpha}{\text{Number of samples that belong to class y} + \alpha K}$$

K is the total number of classes that x_i can take on and α is a constant hyperparameter. Compare this to the formula from before:

$$P(x_i = k|Y = y) = \frac{\text{Number of samples that belong to class y and in which } x_i \text{ takes on class k}}{\text{Number of samples that belong to class y}}$$

Note: Use $\alpha = 1$.
Note: It's important to clarify that "Laplace" smoothing is not related to the Laplace distribution. The

term "Laplace" here simply refers to the technique of smoothing probabilities with pseudocounts.

Note: For the case of Bernoulli distributions, Laplace smoothing is equivalent to using a Beta$(\alpha+1, \alpha+1)$ prior. For categorical distributions, Laplace smoothing can be interpreted as using a Dirichlet prior.

Now, let's walk through an example where our goal is to classify whether an email is SPAM or not. 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:

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

$P(Y=1) = 3/7 \quad P(Y=0) = 4/7$

j	$P(X_j = 1 \mid Y = 1)$	$P(X_j = 1 \mid Y = 0)$
315	1/5	4/6
free	3/5	2/6
is	2/5	1/6
money	4/5	1/6
now	3/5	2/6
Pat	2/5	4/6
Sir	2/5	1/6
teach	2/5	5/6
to	2/5	3/6
tomorrow	1/5	1/6

2. Consider the following email body: X = Pat teach now. Accounting for the laplace smoothing, fill in the table below:

j	x	$P(X_j = x \mid Y = 1)$	$P(X_j = x \mid Y = 0)$
315	0	4/5	2/6
free	0	2/5	4/6
is	0	3/5	5/6
money	0	1/5	5/6
now	1	3/5	2/6
Pat	1	2/5	4/6
Sir	0	3/5	5/6
teach	1	2/5	5/6
to	0	3/5	3/6
tomorrow	0	4/5	5/6

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

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

$P(Y = 1, X_1, ..., X_m) = \prod P(X \mid Y = 1)P(Y = 1)$
4/5 * 2/5 * 3/5 * 1/5 * 3/5 * 2/5 * 3/5 * 2/5 * 3/5 * 4/5 * 3/7
= 0.00045500708

$P(Y = 0, X_1, ..., X_M)$
2/6 * 4/6 * 5/6 * 5/6 * 2/6 * 4/6 * 5/6 * 5/6 * 3/6 * 3/6 * 4/7
=0.00340213817

4.

$P(Y = 1 \mid X_1, ..., X_M)$	$P(Y = 0 \mid X_1, ..., X_M)$
0.00045500708 / (0.00045500708 + 0.00340213817) = 0.118	0.00340213817 / (0.00045500708 + 0.00340213817) = 0.882

We see that $P(Y = 1 \mid X_1, ..., X_M)$ is no longer 0.

What's the effect of α? Compare when $\alpha = 1$ with when $\alpha = \infty$

As α increases, the likelihood probability moves towards uniform distribution.