Lecture #13: Bayes and LDA

CS 109A, STAT 121A, AC 209A: Data Science

Weiwei Pan, Pavlos Protopapas, **Kevin Rader**

Fall 2016

Harvard University

Announcements

- **Midterm**:
  1. Distribution of scores is on Canvas (under announcements).
  2. Generally speaking: we were very pleased!
  3. So what grade did I get?
  4. Solutions are posted under Quizzes.
  5. Questions about MC problems #7-10 send directly to me
  6. Other questions: email the help line

Announcements

- **Midterm**:
  1. Distribution of scores is on Canvas (under announcements).
  2. Generally speaking: we were very pleased!
  3. So what grade did I get?
  4. Solutions are posted under Quizzes.
  5. Questions about MC problems #7-10 send directly to me
  6. Other questions: email the help line
- **Walkout!**

Quiz Time

Time for Quiz...password is **joeexotic2016**

Lecture Outline

Bayes' Theorem

Diagnostic Testing

ROC Curves

Linear Discriminant Analysis (LDA)

Bayes' Theorem

We defined conditional probability as:

$$P(B|A) = \frac{P(B \cap A)}{P(A)}$$

And using the fact that $P(B \cap A) = P(A|B)P(B)$ we get Bayes' Theroem:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Another version of Bayes' Theorem is found by substituting in the Law of Total Probability (LOTP) into the denominator:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^C)P(B^C)}$$

Where have we seen Bayes' Theorem before? Why do we care?

Lecture Outline

Bayes' Theorem

Diagnostic Testing

ROC Curves

Linear Discriminant Analysis (LDA)

Diagnostic Testing

In the diagnostic testing paradigm, one cares about whether the results of a test (like a classification test) matches truth (the true class that observation belongs to). The simplest version of this is trying to detect disease ($D+$ vs. $D-$) based on a diagnostic test ($T+$ vs. $T-$).

Medical examples of this include various screening tests: breast cancer screening through (i) self-examination and (ii) mammographies, prostate cancer screening through (iii) PSA tests, and Colo-rectal cancer through (iv) colonoscopies.

These tests are a little controversial because of poor predictive probability of the tests.

Diagnostic Testing (cont.)

Bayes' theorem can be rewritten for diagnostic tests:

$$P(D+|T+) = \frac{P(T+|D+)P(D+)}{P(T+|D+)P(D+) + P(T-|D-)P(D-)}$$

These probability quantities can then be defined as:

- *Sensitivity*: $P(T+|D+)$
- *Specificity*: $P(T-|D-)$
- *Prevalence*: $P(D+)$
- *Positive Predictive Value*: $P(D+|T+)$
- *Negative Predictive Value*: $P(D-|T-)$

How do positive and negative predictive value relate? Be careful...

Diagnostic Testing (cont.)

We mentioned that these tests are a little controversial because of their poor predictive probability. When will these tests have poor positive predictive probability?

Diagnostic Testing (cont.)

We mentioned that these tests are a little controversial because of their poor predictive probability. When will these tests have poor positive predictive probability?

When the disease is not very prevalent, then the number of 'false positives' will overwhelm the number of true positive. For example, PSA screening for prostate cancer has sensitivity of about 90% and specificity of about 97% for some age groups (men in their fifties), but prevalence is about 0.1%.

What is positive predictive probability for this diagnostic test?

Why do we care?

As data scientists, why do we care about diagnostic testing from the medical world? (hint: it's not just because Kevin is a trained biostatistician!)

Why do we care?

As data scientists, why do we care about diagnostic testing from the medical world? (hint: it's not just because Kevin is a trained biostatistician!)

Because classification can be thought of as a diagnostic test. Let $Y_i = k$ be the event that observation i truly belongs to category k, and let $\hat{Y}_i = k$ be the event that we correctly predict it to be in class k. Then Bayes' rule states that our *Positive Predictive Value* for classification is:

$$P(Y_i = k|\hat{Y}_i = k) = \frac{P(\hat{Y}_i = k|Y_i = k)P(Y_i = k)}{P(\hat{Y}_i = k|Y_i = k)P(Y_i = k) + P(\hat{Y}_i = k|Y_i \neq k)P(Y_i \neq k)}$$

Thus the probability of a predicted outcome truly being in a specific group depends on what?

Why do we care?

As data scientists, why do we care about diagnostic testing from the medical world? (hint: it's not just because Kevin is a trained biostatistician!)

Because classification can be thought of as a diagnostic test. Let $Y_i = k$ be the event that observation i truly belongs to category k, and let $\hat{Y}_i = k$ be the event that we correctly predict it to be in class k. Then Bayes' rule states that our *Positive Predictive Value* for classification is:

$$P(Y_i = k|\hat{Y}_i = k) = \frac{P(\hat{Y}_i = k|Y_i = k)P(Y_i = k)}{P(\hat{Y}_i = k|Y_i = k)P(Y_i = k) + P(\hat{Y}_i = k|Y_i \neq k)P(Y_i \neq k)}$$

Thus the probability of a predicted outcome truly being in a specific group depends on what? The proportion of observations in that class!

Error in Classification

There are 2 major types of error in classification problems based on a binary outcome. They are:

- False positives: incorrectly predicting $\hat{Y} = 1$ when it truly is in $Y = 0$.
- False negative: incorrectly predicting $\hat{Y} = 0$ when it truly is in $Y = 1$.

The results of a classification algorithm are often summarized in two ways: a confusion table, sometimes called a contingency table, or a 2x2 table (more generally kxk table) and an receiver operating characteristics (ROC) curve.

Confusion table

When a classification algorithm (like logistic regression) is used, the results can be summarize in a kxk table as such:

		True Republican Status	
		Yes	No
Predicted	Yes	487	288
Republican	No	218	314

The table above was a classification based on a logistic regression model to predict political party (Dem. vs. Rep.) based on 3 predictors: $X_1 =$ whether respondent believes abortion is legal, $X_1 =$ income (logged) and $X_3 =$ years of education.

What are the false positive and false negative rates for this classifier?

Bayes' Classifier Choice

A classifier's error rates can be tuned to modify this table. How?

Bayes' Classifier Choice

A classifier's error rates can be tuned to modify this table. How?

The choice of the Bayes' classifier level will modify the characteristics of this table.

If we thought is was more important to predict republicans correctly (lower false positive rate), what could we do for our Bayes' classifier level?

We could classify instead based on:

$$\hat{P}(Y = 1) < \pi$$

and we could choose π to be some level other than 0.5. Let's see what the table looks like if π were 0.28 or 0.52 instead (why such strange numbers?).

Other Confusion table

Based on $\pi = 0.28$:

		True Republican Status	
		Yes	No
Predicted	Yes	247	528
Republican	No	80	452

What has improved? What has worsened?

Other Confusion table

Based on $\pi = 0.28$:

		True Republican Status	
		Yes	No
Predicted	Yes	247	528
Republican	No	80	452

What has improved? What has worsened?Based on $\pi = 0.52$:

		True Republican Status	
		Yes	No
Predicted	Yes	627	148
Republican	No	388	144

Which should we choose? Why?

Lecture Outline

Bayes' Theorem

Diagnostic Testing

ROC Curves

Linear Discriminant Analysis (LDA)

ROC Curves

The ROC curve illustrates the trade-off for all possible thresholds chosen for the two types of error (or correct classification).

The vertical axis displays the true positive predictive value and the horizontal axis depicts the true negative predictive value.

What is the shape of an ideal ROC curve?

ROC Curves

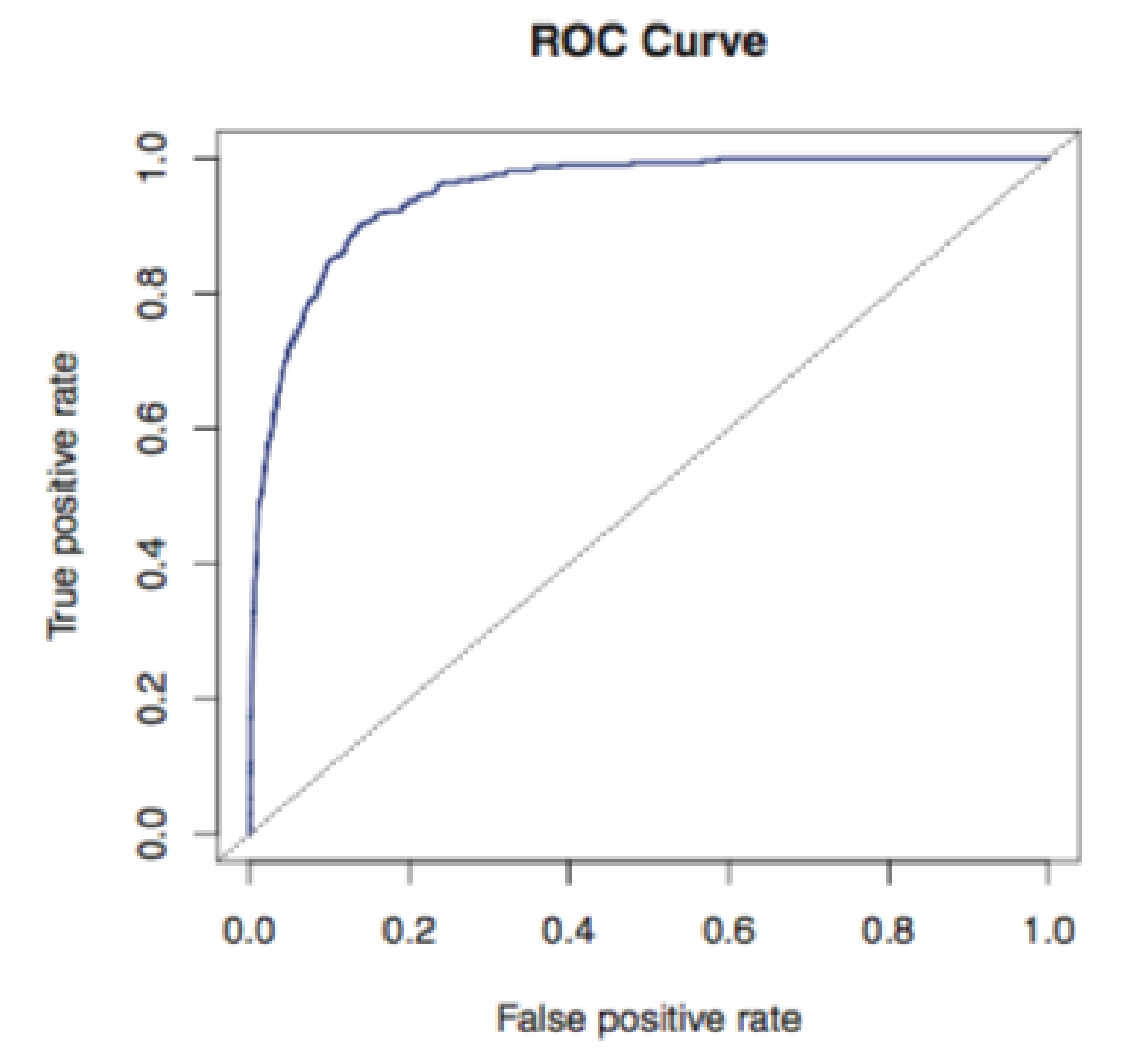
The ROC curve illustrates the trade-off for all possible thresholds chosen for the two types of error (or correct classification).

The vertical axis displays the true positive predictive value and the horizontal axis depicts the true negative predictive value.

What is the shape of an ideal ROC curve?

See next slide for an example.

ROC Curve Example



ROC Curve for measuring classifier preformance

The overall performance of a classifier, calculated over all possible thresholds, is given by the area under the ROC curve ('AUC').

An ideal ROC curve will hug the top left corner, so the larger the AUC the better the classifier.

What is the worst case scenario for AUC? What is the best case? What is AUC if we independently just flip a coin to perform classification?

ROC Curve for measuring classifier preformance

The overall performance of a classifier, calculated over all possible thresholds, is given by the area under the ROC curve ('AUC').

An ideal ROC curve will hug the top left corner, so the larger the AUC the better the classifier.

What is the worst case scenario for AUC? What is the best case? What is AUC if we independently just flip a coin to perform classification?

This AUC then can be use to compare various approaches to classification: Logistic regression, LDA (to come), kNN, etc...

Lecture Outline

Bayes' Theorem

Diagnostic Testing

ROC Curves

Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA)

Linear discriminant analysis (LDA) takes a different approach to classification than logistic regression. Rather than attempting to model the conditional distribution of Y given X, $P(Y = k|X = x)$, LDA models the distribution of the predictors X given the different categories that Y takes on, $P(X = x|Y = k)$. In order to flip these distributions around to model $P(X = x|Y = k)$ an analyst uses Bayes' theorem.

In this setting with one feature (one X), Bayes' theorem can then be written as:

$$P(Y = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{j=1}^{K} f_j(x)\pi_j}$$

What does this mean?

Linear Discriminant Analysis (LDA)

Linear discriminant analysis (LDA) takes a different approach to classification than logistic regression. Rather than attempting to model the conditional distribution of Y given X, $P(Y = k|X = x)$, LDA models the distribution of the predictors X given the different categories that Y takes on, $P(X = x|Y = k)$. In order to flip these distributions around to model $P(X = x|Y = k)$ an analyst uses Bayes' theorem.

In this setting with one feature (one X), Bayes' theorem can then be written as:

$$P(Y = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{j=1}^{K} f_j(x)\pi_j}$$

What does this mean?

Linear Discriminant Analysis (LDA)

$$P(Y = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{j=1}^{K} f_j(x)\pi_j}$$

The left hand side, $P(Y = k|X = x)$, is called the *posterior* probability and gives the probability that the observation is in the k^{th} category given the feature, X, takes on a specific value, x. The numerator on the right is conditional distribution of the feature within category k, $f_k(x)$, times the *prior* probability that observation is in the k^{th} category.

The *Bayes' classifier* is then selected. That is the observation assigned to the group for which the posterior probability is the largest.

LDA for one predictor

LDA has the simplest form when there is just one predictor/feature ($p = 1$). In order to estimate $f_k(x)$, we have to assume it comes from a specific distribution. If X is quantitative, what distribution do you think we should use?

LDA for one predictor

LDA has the simplest form when there is just one predictor/feature ($p = 1$). In order to estimate $f_k(x)$, we have to assume it comes from a specific distribution. If X is quantitative, what distribution do you think we should use?

One common assumption is that $f_k(x)$ comes from a Normal distribution:

$$f_k(x) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right).$$

In shorthand notation, this is often written as $X|Y = k \sim N(\mu_k, \sigma_k^2)$, meaning, the distribution of the feature X within category k is Normally distributed with mean μ_k and variance σ_k^2.

LDA for one predictor (cont.)

An extra assumption that the variances are equal, $\sigma_1^2 = \sigma_2^2 = ... = \sigma_K^2$ will simplify are lives.

Plugging this assumed likelihood into the Bayes' formula (to get the posterior) results in:

$$P(Y = k|X = x) = \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu_k)^2}{2\sigma^2}\right)}{\sum_{j=1}^{K} \pi_j \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu_j)^2}{2\sigma^2}\right)}$$

The Bayes classifier will be the one that maximizes this over all values chosen for x. How should we maximize?

So we take the log of this expression and rearrange to simplify our maximization...

LDA for one predictor (cont.)

So we maximize the following simplified expression:

$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k$$

How does this simplify if we have just two classes ($K = 2$) and if we set our prior probabilities to be equal?

LDA for one predictor (cont.)

So we maximize the following simplified expression:

$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k$$

How does this simplify if we have just two classes ($K = 2$) and if we set our prior probabilities to be equal? This is equivalent to choosing a decision boundary for x for which

$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}$$

Intuitively, why does this expression make sense? What do we use in practice?

LDA for one predictor (cont.)

In practice we do ont know the true mean, variance, and prior. So we estimate them with the classical estimates, and plug-them into the expression:

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

and

$$\hat{\sigma}_k^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$

where n is the total sample size and n_k is the sample size within class k (thus, $n = \sum n_k$).

LDA for one predictor (cont.)

This classifier works great if the classes are about equal in proportion, but can easily be extended to unequal class sizes. Instead of assuming all priors are equal, we instead set the priors to match the 'prevalence' in the data set:

$$\hat{\pi}_k = \hat{n}_k / n$$

Note: we can use a prior probability from knowledge of the subject as well; for example, if we expect the test set to have a different prevalence than the training set. How could we do this in the Dem. vs. Rep. data set?

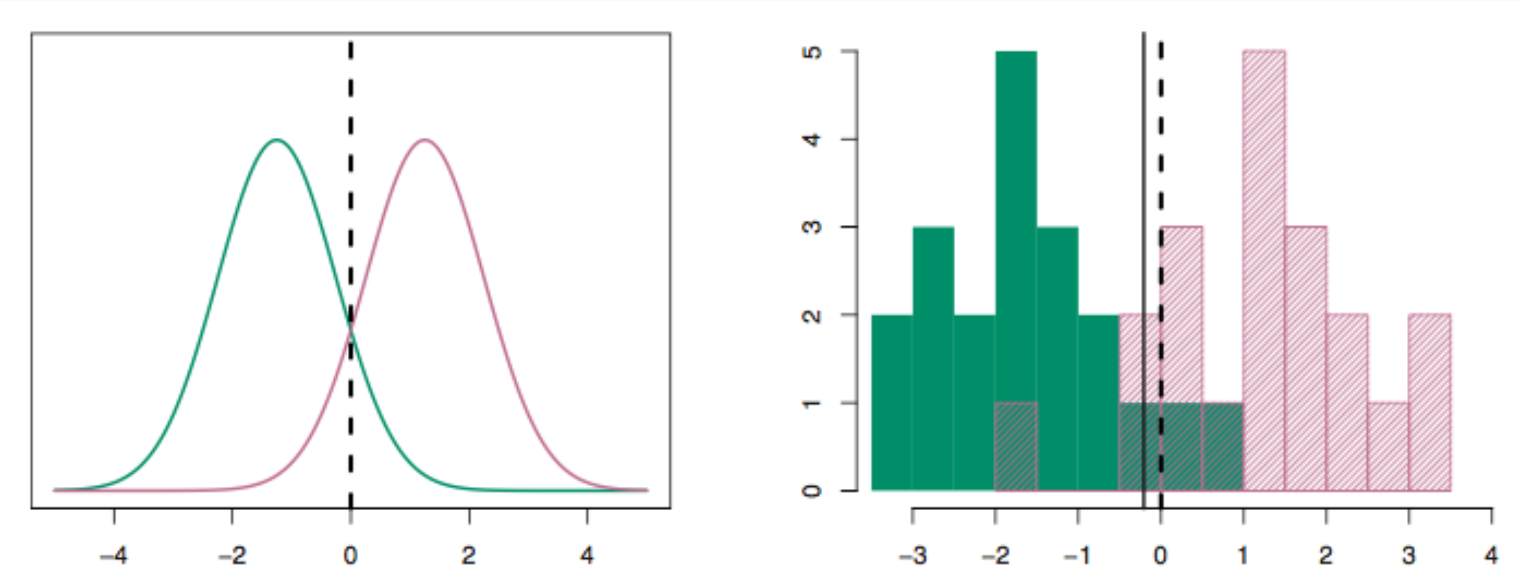
LDA for one predictor (cont.)

Plugging all of these estimates back into the original logged maximization formula we get:

$$\hat{\delta}_k(x) = x\frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log \hat{\pi}_k$$

Thus this classifier is called the linear discriminant classifier: this discriminant function is a linear function of x.

Illustration of LDA when $p = 1$



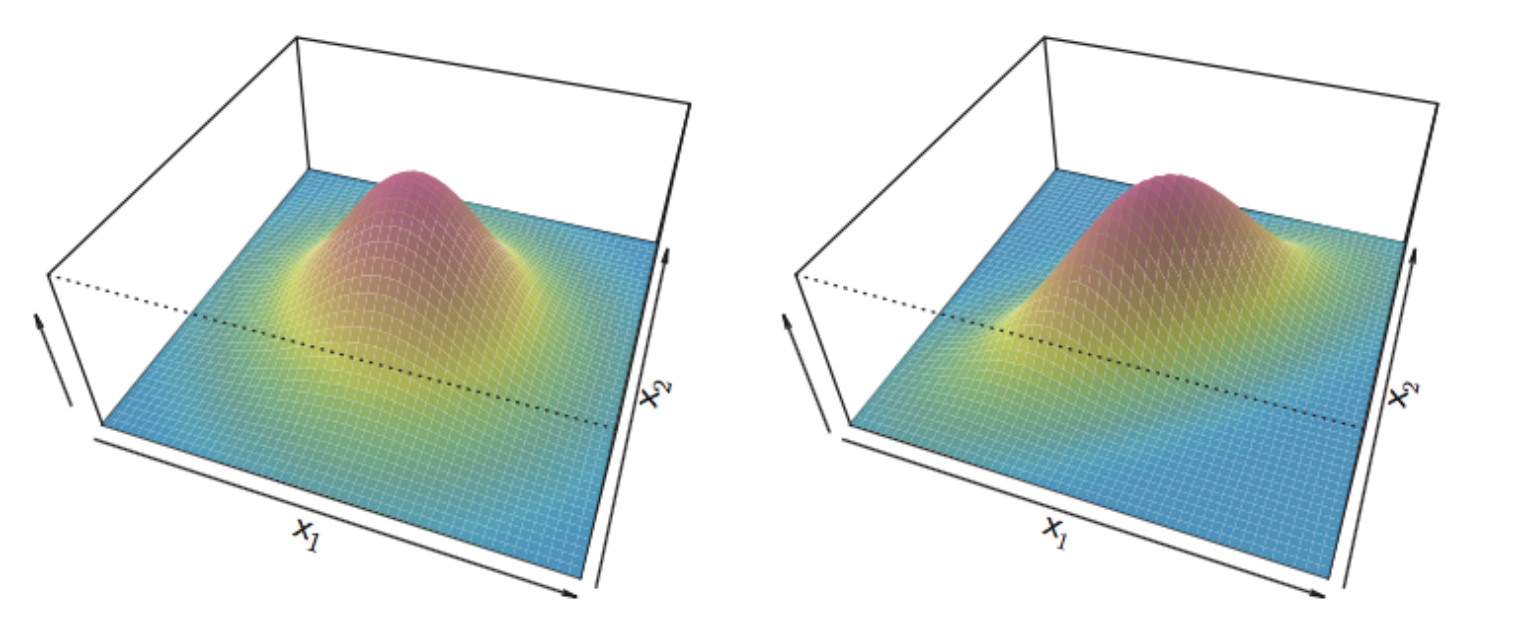
LDA when $p > 1$

LDA generalizes 'nicely' to the case when there is more than one predictor. Instead of assuming the one predictor is Normally distributed, it assumes that the set of predictors for each class it 'multivariate normal distributed' (shorthand: MVN). What does that mean?

LDA when $p > 1$

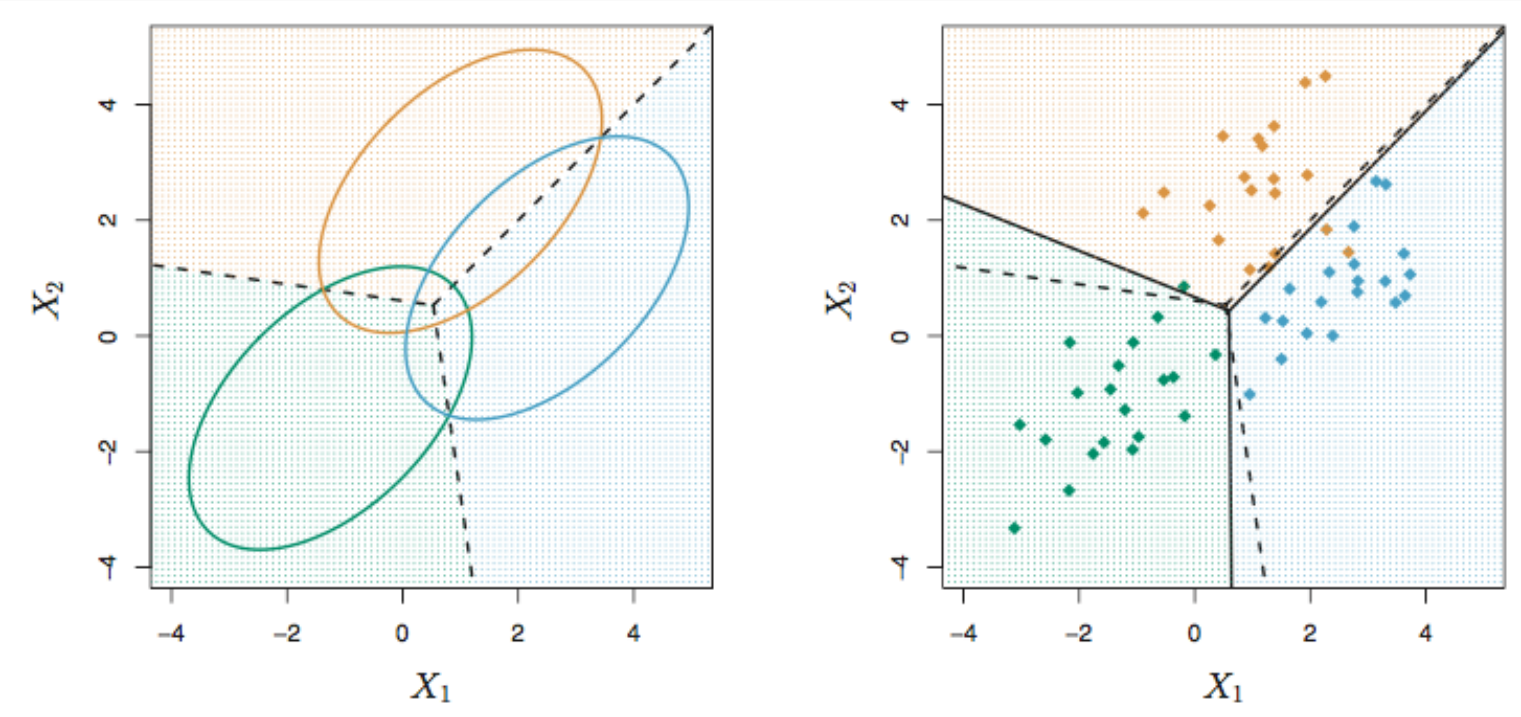
LDA generalizes 'nicely' to the case when there is more than one predictor. Instead of assuming the one predictor is Normally distributed, it assumes that the set of predictors for each class it 'multivariate normal distributed' (shorthand: MVN). What does that mean? This means that the vector of X for an observation has a multidimensional normal distribution with a mean vector, μ, and a covariance matrix, Σ. We'll get more into this in the next lecture, but a picture is worth a thousand words.

MVN distribution for 2 variables



LDA when $p > 1$

The linear discrimnation nature of LDA still holds when $p > 1$ (and when $K > 2$ for that matter as well). A picture can be very illustrative:



Inventor of LDA: R.A. Fisher

The 'Father' of Statistics. More famous for work in genetics (statistically concluded that Mendel's genetic experiments were 'massaged'). Novel statistical work includes:

1. Experimental Design
2. ANOVA
3. F-test (why do you think it's called the F-test?)
4. Exact test for 2x2 tables
5. Maximum Likelihood Theory
6. Use of $\alpha = 0.05$ significance level: "The value for which P = .05, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not."
7. And so much more...