STEVENS INSTITUTE of TECHNOLOGY
School of Business

BIA-652

Multivariate Data Analytics

Classification:
Discriminant Analysis and Logistic Regression

Prof. Feng Mai
School of Business

For academic use only.

Case 1: Brand Preference for Orange Juice

We would like to predict if customers prefer to buy Citrus Hill Orange Juice

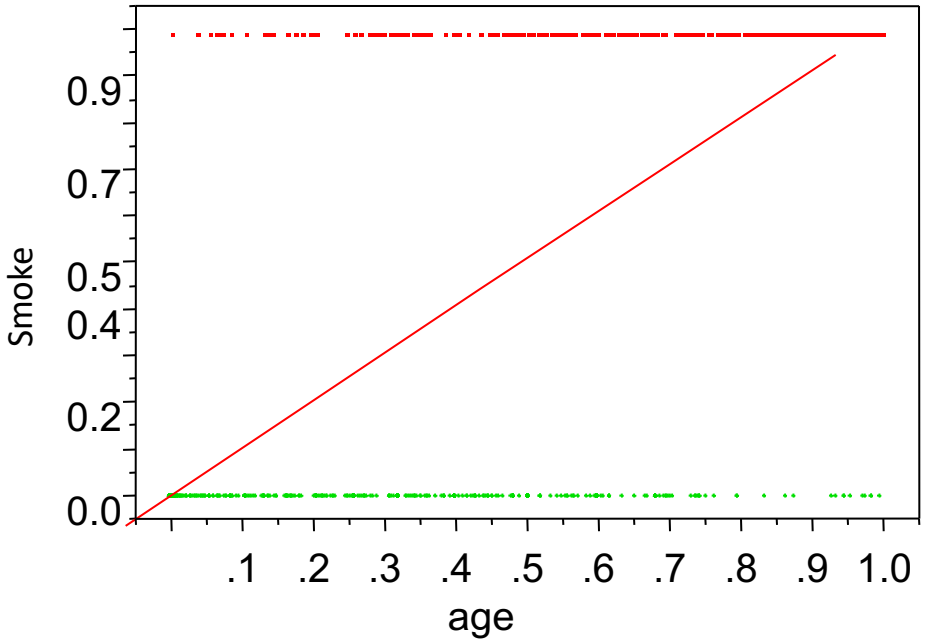
The Y (Purchase) variable is categorical: 0 or 1

The X (LoyalCH) variable is a numerical value (between 0 and 1) which specifies the how much the customers are loyal to the Citrus Hill (CH) orange juice

Can we use Linear Regression when Y is categorical?

Why not Linear Regression?

- When Y only takes on values of 0 and 1, why standard linear regression in inappropriate?



How do we interpret values smaller than 1?

How do we interpret values of Y between 0 and 1?

Problems

The regression line $\beta_0+\beta_1 X$ can take on any value between negative and positive infinity

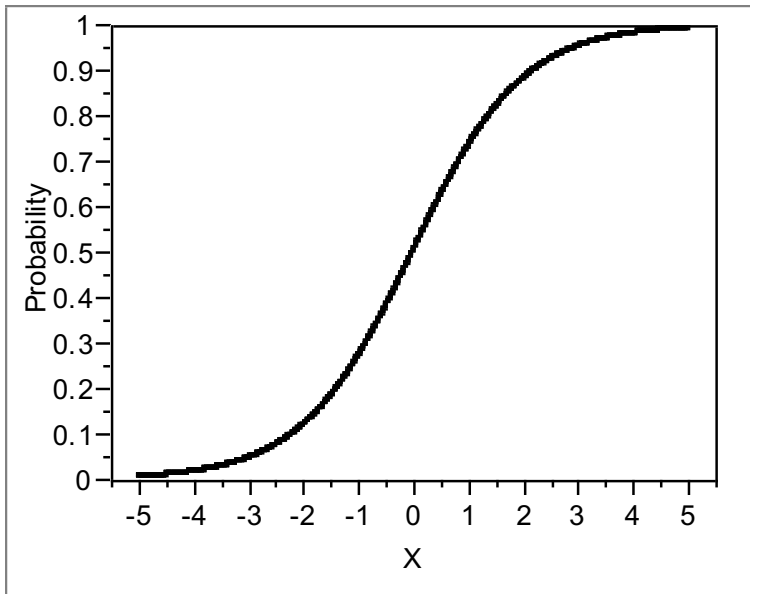
- In the orange juice classification problem, Y can only take on two possible values: 0 or 1.

- Therefore the regression line almost always predicts the wrong value for Y in classification problems

Solution: Use Logistic Function

- Instead of trying to predict Y, let's try to predict P(Y = 1), i.e., the probability a customer buys Citrus Hill (CH) juice.
- Thus, we can model P(Y = 1) using a function that gives outputs between 0 and 1.
- We can use the logistic function to "squash" the output to [0,1]
- Logistic Regression
- $P(Y = 1) = \frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x_1}}$
- $\ln\left(\frac{P(Y=1)}{P(Y=0)}\right) = \beta_0 + \beta_1 x_1$

  log odds



Fitting Logistic Regression

- Similar to linear regression, we need to estimate the β parameters using data
- Maximum likelihood estimation
  - The data generating process is binomial
  - What are the β parameters that maximizes the probability of generating the observed data?

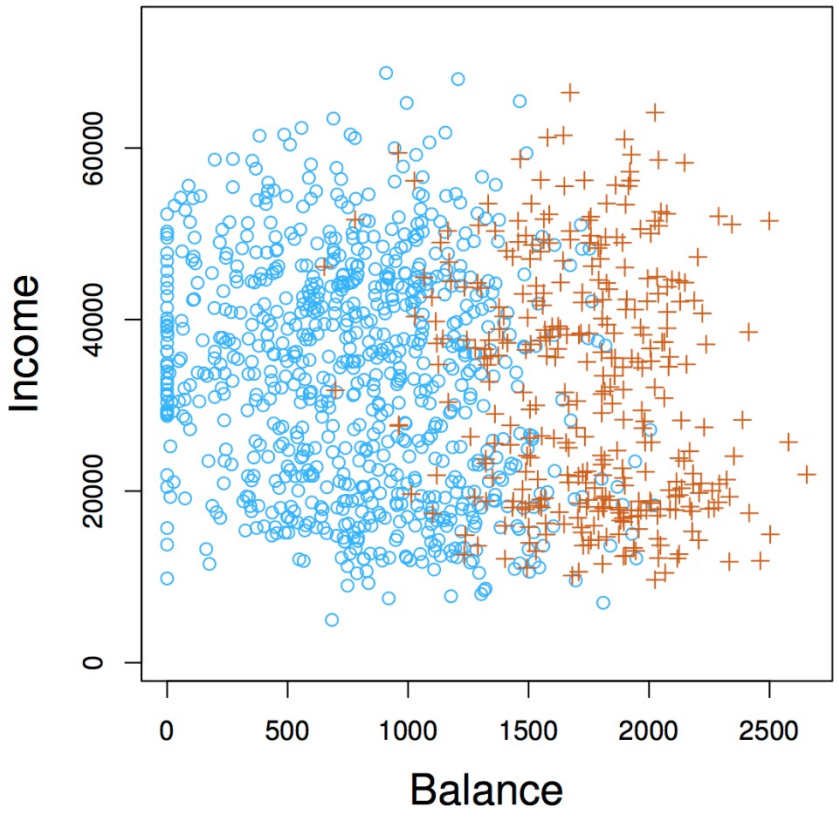
Max $$L(\beta_0, \beta) = \prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i)^{1-y_i}$$

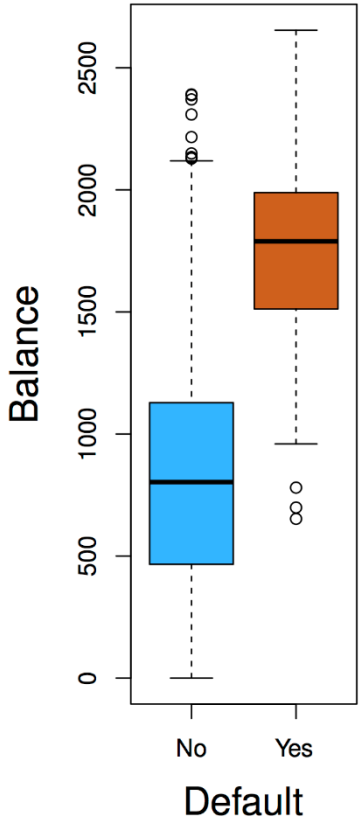
$$L(\beta; \mathbf{y}, \mathbf{X}) = \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{X}_i\beta)}{1 + \exp(\mathbf{X}_i\beta)} \right)^{y_i} \left(\frac{1}{1 + \exp(\mathbf{X}_i\beta)} \right)^{1-y_i}.$$

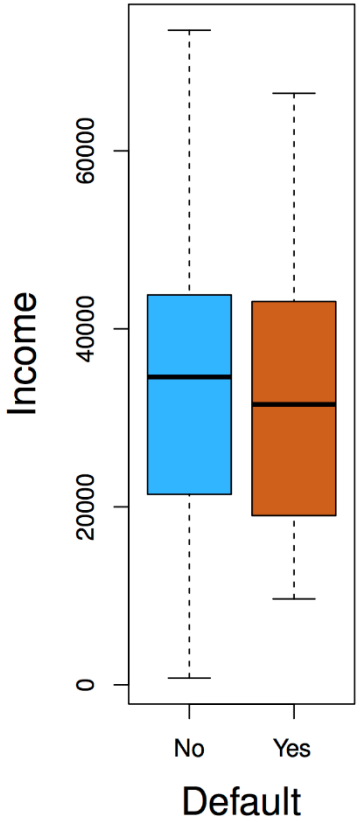
- No closed-form solution, needs numerical optimization. See reading for details.

Case 2: Credit Card Default Data

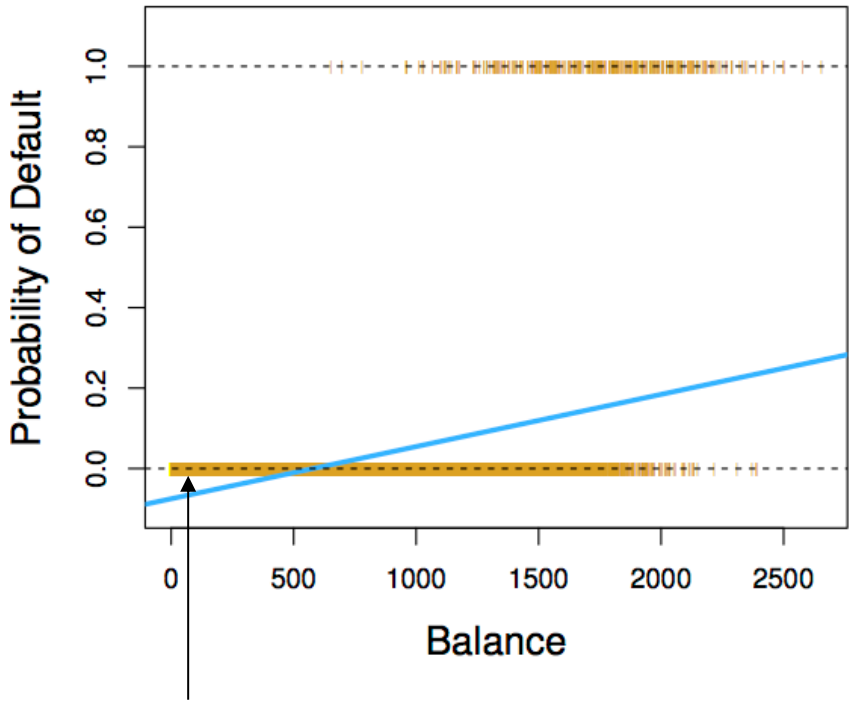
- We would like to be able to predict customers that are likely to default
- Possible X variables are:
  - Annual Income
  - Monthly credit card balance
- The Y variable (Default) is categorical: Yes or No
- How do we check the relationship between Y and X?







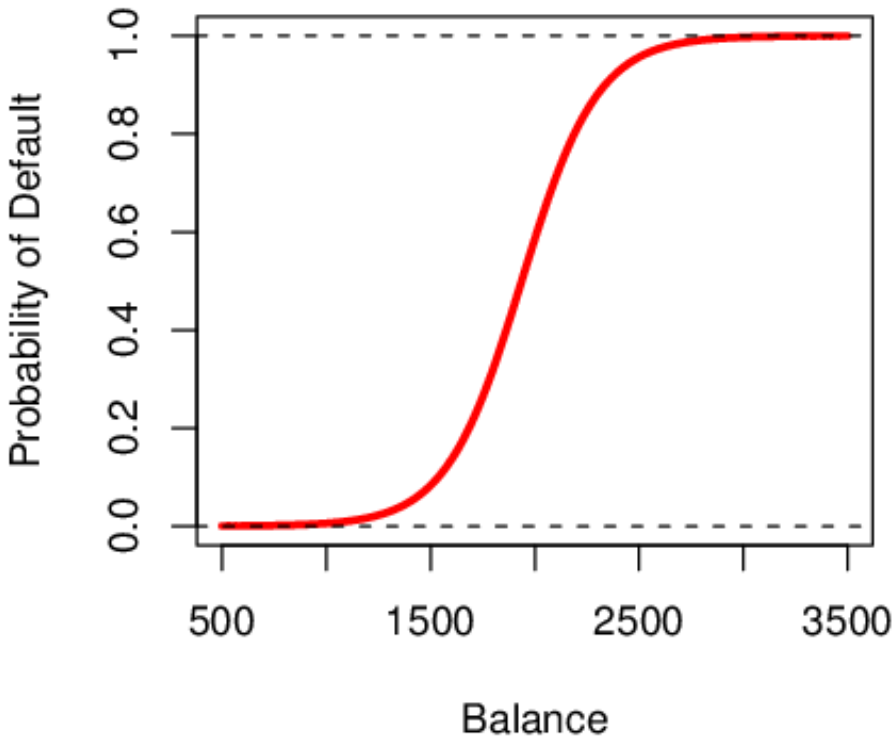
- If we fit a linear regression to the Default data, then for very low balances we predict a negative probability, and for high balances we predict a probability above 1!



When Balance < 500,

Pr(default) is negative!

- Now the probability of default is close to, but not less than zero for low balances. And close to but not above 1 for high balances



Interpreting β

- We are predicting P(Y) and not Y.
- If β_1 =0, this means that there is no relationship between Y and X.
- If β_1 >0, this means that when X gets larger, so does the probability that Y = 1.
- If β_1 <0, this means that when X gets larger, the probability that Y = 1 gets smaller.

Interpreting β

- Logistic regression coefficients tell us about log *odds ratios*
- Example: "Among 80 year-old men, the odds of being dead before age 90 are three times as great for smokers."

$$\frac{\text{Odds of death given smoker}}{\text{Odds of death given nonsmoker}} = 3$$

- X=1 means smoker, X=0 means non-smoker
- Y=1 means dead, Y=0 means alive
- Log odds of death = $\beta_0 + \beta_1 x$
- Odds of death = $e^{\beta_0} e^{\beta_1 x}$

β_1 = 1.1

Group	x	Odds of Death
Smokers	1	$e^{\beta_0} e^{\beta_1}$
Non-smokers	0	e^{β_0}

$$\frac{\text{Odds of death given smoker}}{\text{Odds of death given nonsmoker}} = \frac{e^{\beta_0} e^{\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

Interpreting β

- If β_1 = 1.1 and statistically significant, exp(1.1) = 3
- Can we simply say: “Smokers who are 80 years-old are 3 times more likely being dead before age 90 compared with non-smokers”?
  - That is, can we say
  - P(Dead | Smoke) / P (Dead | Non-Smoke) = 3 ?
- It depends!
- (Odds ≈ Probability only when Probability is small)
- Excel Demo

Are the coefficients significant?

- We still want to perform a hypothesis test to see whether we can be sure that are β_0 and β_1 significantly different from zero.
- We use a Z test instead of a T test, but of course that doesn't change the way we interpret the p-value
- Here the p-value for balance is very small, and b_1 is positive, so we are sure that if the balance increase, then the probability of default will increase as well.

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.6513	0.3612	-29.5	< 0.0001
balance	0.0055	0.0002	24.9	< 0.0001

Making Prediction

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.6513	0.3612	-29.5	< 0.0001
balance	0.0055	0.0002	24.9	< 0.0001

- Suppose an individual has an average balance of $1000. What is their probability of default?

- The predicted probability of default for an individual with a balance of $1000 is less than 1%.

- For a balance of $2000, the probability is much higher, and equals to 0.586 (58.6%).

Categorical Predictors in Logistic Regression

- We can predict if an individual default by checking if she is a student or not. Thus we can use a categorical variable “Student” coded as (Student = 1, Non-student =0).
- b_1 is positive: This indicates students tend to have higher default probabilities than non-students

	Coefficient	Std. Error	Z-statistic	P-value
`Intercept`	-3.5041	0.0707	-49.55	< 0.0001
`student[Yes]`	0.4049	0.1150	3.52	0.0004

Multiple Logistic Regression- Default Data

Predict Default using:

- Balance (quantitative)
- Income (quantitative)
- Student (dummy)

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.8690	0.4923	-22.08	< 0.0001
balance	0.0057	0.0002	24.74	< 0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	-0.6468	0.2362	-2.74	0.0062

Predictions

- A student with a credit card balance of $1,500 and an income of $40,000 has an estimated probability of default

$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times1500+0.003\times40-0.6468\times1}}{1+e^{-10.869+0.00574\times1500+0.003\times40-0.6468\times1}} = 0.058.$$

Example 2: Effect of Workplace Smoking Ban

- The dependent variable takes two values, coded as 0 and 1
  - Policy Evaluation: Do workplace smoking bans reduce smoking (controlling for age, gender, etc.)?
    - Smoker = yes → 1
    - Smoker = no → 0

smoker	ban	age	education	afam	hispanic	gender
yes	yes	41	hs	no	no	female
yes	yes	44	some college	no	no	female
no	no	19	some college	no	no	female
yes	no	29	hs	no	no	female
no	yes	28	some college	no	no	female
no	no	40	some college	no	no	male
yes	yes	47	some college	no	no	female

Example 2: Interpretation of Coefficients

EXP(b) provides the odds ratio (OR)

- odds ratio > 1 : An increase in X is associated with an **increase** in P(Y = 1)
- odds ratio < 1 : An increase in X is associated with a **decrease** in P(Y = 1)
- b > 0 → EXP (b) > 1
- b < 0 → EXP (b) > 1

Example:

Working Place Smoking Ban: b = -0.251, p < 0.01

OR = EXP (-0.251) = 0.778

Controlling for other factors, workplace smoking ban is associated with a reduced odds of smoking (OR = 0.778, P < 0.01).

- A 22.2% reduction in the odds of smoking

	Dependent variable:
	smoker
Age	-0.007***
	(0.002)
Workplace Smoking Ban	-0.251***
	(0.049)
Female	-0.189***
	(0.049)
High school	-0.408***
	(0.083)
Some college	-0.751***
	(0.087)
College	-1.506***
	(0.101)
Master	-1.931***
	(0.131)
African-American	-0.149*
	(0.090)
Hispanic	-0.585***
	(0.083)
Constant	0.234**
	(0.116)
Observations	10,000
Log Likelihood	-5,251.095
Akaike Inf. Crit.	10,522.190
Note:	*p<0.1; **p<0.05; ***p<0.01

Linear Discriminant Analysis (LDA)

*** Not to be confused with Latent Dirichlet Allocation (LDA)**

Linear Discriminant Analysis (LDA)

- LDA undertakes the same task as Logistic Regression. It classifies data based on a categorical dependent variable
- LDA involves the determination of linear equation (just like linear regression) that will predict which group the case belongs to.

$$D = v_1 X_1 + v_2 X_2 + \ldots + v_i X_i + a$$

  - D: discriminant function
  - v: discriminant coefficient or weight for the variable
  - X: variable
  - a: constant

Altman's Z Score

Z-Score = 1.2A + 1.4B + 3.3C + 0.6D + 1.0E

Where:

- A = working capital / total assets
- B = retained earnings / total assets
- C = earnings before interest and tax / total assets
- D = market value of equity / total liabilities
- E = sales / total assets

- A score below 1.8 means it's likely the company is headed for bankruptcy, while companies with scores above 3 are not likely to go bankrupt. Investors can use Altman Z-scores to determine whether they should buy or sell a stock if they're concerned about the company's underlying financial strength.

- Source: https://www.investopedia.com/terms/a/altman.asp

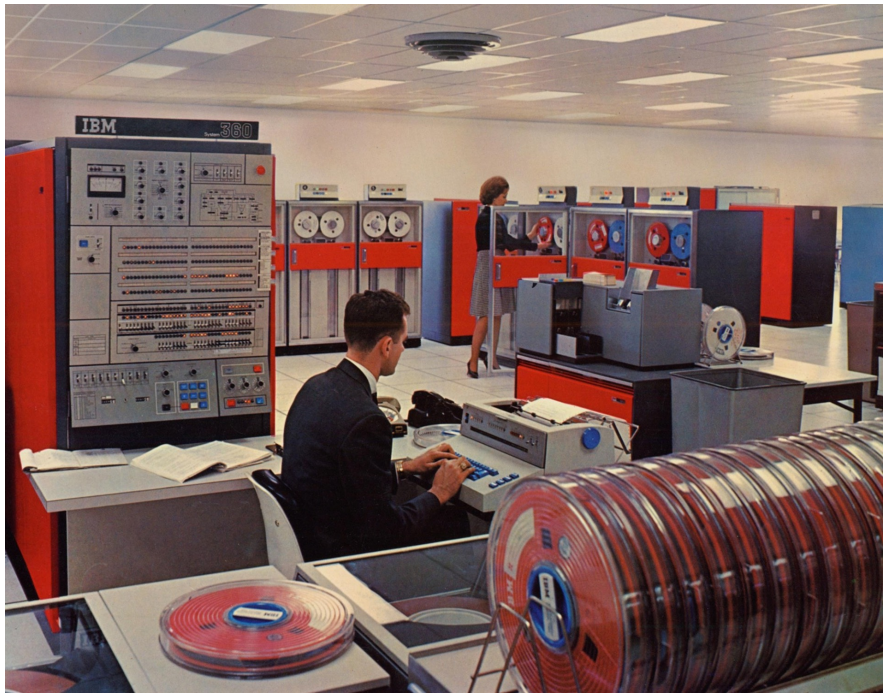
Logistic Regression (LR) vs LDA

- In the case where n is small, and the distribution of predictors X is approximately normal, or the classes are well separated, then LDA is more stable than LR
- They are both linear classifiers, so empirically they often provide similar results
- Coefficients of LR are easier to interpret (Log Odds Ratio)
- LDA can be computed by hand (hence important for historical reasons), but LR is more popular nowadays
- LDA is generative, LR is discriminative (see this post)

Generative vs Discriminative

Generative classifiers learn a model of the joint probability, $p(x, y)$, of the inputs x and the label y, and make their predictions by using Bayes rules to calculate $p(y|x)$, and then picking the most likely label y. Discriminative classifiers model the posterior $p(y|x)$ directly, or learn a direct map from inputs x to the class labels. There are several compelling reasons for using discriminative rather than generative classifiers, one of which, succinctly articulated by Vapnik [6], is that "one should solve the [classification] problem directly and never solve a more general problem as an intermediate step [such as modeling $p(x|y)$]." Indeed, leaving aside computational issues and matters such as handling missing data, the prevailing consensus seems to be that discriminative classifiers are almost always to be preferred to generative ones.

Ng & Jordan (2002)



Victor 3900 engineering prototype

Victor 3900 desktop calculator

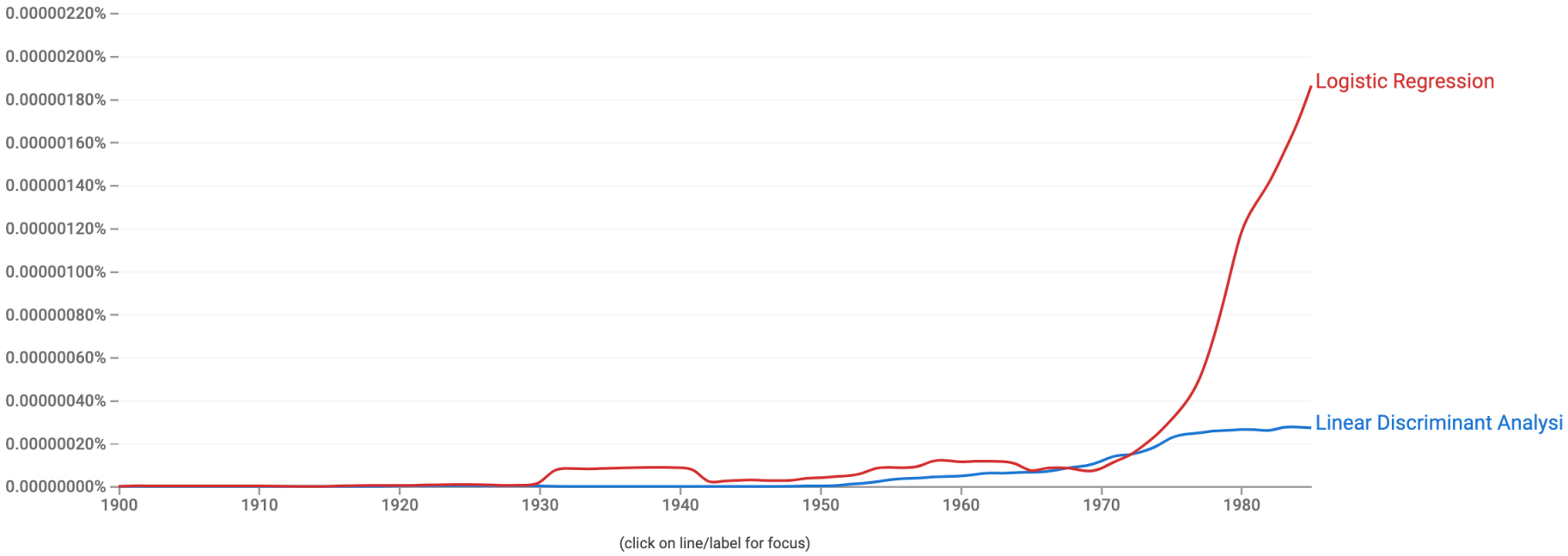
- *Memory & Storage*

Victor Comptometer Corporation produces the Victor 3900 desktop calculator. Six 100-bit MOS shift registers built by General Microelectronics provided memory for the calculator, which was the first to use MOS for both logic and memory. The calculator could perform multiple functions and had a small, integrated CRT display. However, the immature MOS manufacturing process made the parts unreliable, limiting sales.

Google Books Ngram Viewer

Linear Discriminant Analysis,Logistic Regression

1900 - 1985 | English (2019) | Case-Insensitive | Smoothing of 5



LDA from Bayes' Theorem (single X)

- With Logistic Regression we modeled the probability of Y being from the k[th] class as

$$\Pr(Y = k|X = x) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}.$$

- However, Bayes' Theorem states

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}.$$

$\pi_k = \Pr(Y = k)$: Probability of any point coming from class k (prior probability)

$f_k(x) = \Pr(X = x|Y = k)$: Density function for X given that X is an observation from class k

Estimate π_k and $f_k(x)$

- We can estimate π_k and $f_k(x)$ to compute $p(X)$
- The most common model for $f_k(x)$ is the Normal Density

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

- Using the density, we only need to estimate three quantities to compute :

$$\mu_k \qquad \sigma_k^2 \qquad \pi_k$$

- To simplify, we can assume that σ^2 is the same for all *K* classes

Use Training Data for Estimation

- The mean μ_k could be estimated by the average of all training observations from the k[th] class.
- The variance σ_k^2 could be estimated as the weighted average of variances of all k classes.
- And, π_k is estimated as the proportion of the training observations that belong to the k[th] class.

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

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

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

where $\hat{\sigma}_k^2 = \frac{1}{n_k-1} \sum_{i:\, y_i=k} (x_i - \hat{\mu}_k)^2$ is the usual formula for the estimated variance in the kth class.

Bayes Theorem:

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}. \qquad (4.10)$$

Normal Density (with estimated parameters from data)

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right), \qquad (4.11)$$

Plug 4.11 into 4.10

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_l)^2\right)}. \qquad (4.12)$$

Take log and simplify

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k) \qquad (4.13)$$

Simple case: 2 classes (K = 2), and $\pi_1 = \pi_2$:

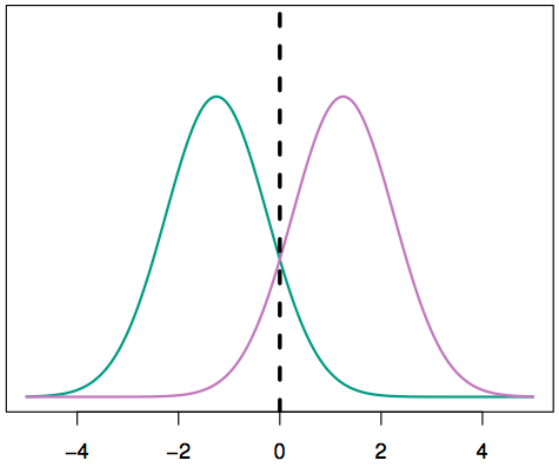
A Simple Example with One Predictor (p =1)

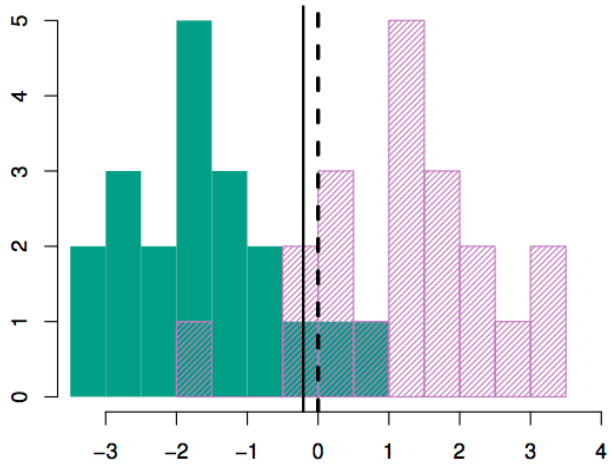
- Suppose we have only one predictor (p = 1)
- Two normal density function $f_1(x)$ and $f_2(x)$, represent two distinct classes
- The two density functions overlap, so there is some uncertainty about the class to which an observation with an unknown class belongs
- The dashed vertical line represents Bayes' decision boundary

Apply LDA

- LDA starts by assuming that each class has a normal distribution with a common variance
- The mean and the variance are estimated
- Finally, Bayes' theorem is used to compute p_k and the observation is assigned to the class with the maximum probability among all k probabilities

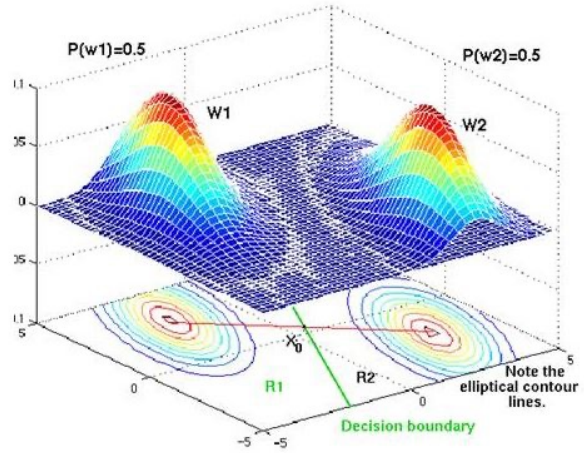
- 20 observations were drawn from each of the two classes
- The dashed vertical line is the Bayes' decision boundary
- The solid vertical line is the LDA decision boundary
  - Bayes' error rate: 10.6%
    - Bayes error rate is the lowest possible error rate for any classifier.
  - LDA error rate: 11.1%
- Thus, LDA is performing pretty well!





When p > 1

- If X is multidimensional (p > 1), we use the same approach except the density function *f(x)* is modeled using the multivariate normal density



$$\vec{w} = \Sigma^{-1}(\vec{\mu}_1 - \vec{\mu}_0)$$

$$c = \vec{w} \cdot \frac{1}{2}(\vec{\mu}_1 + \vec{\mu}_0)$$

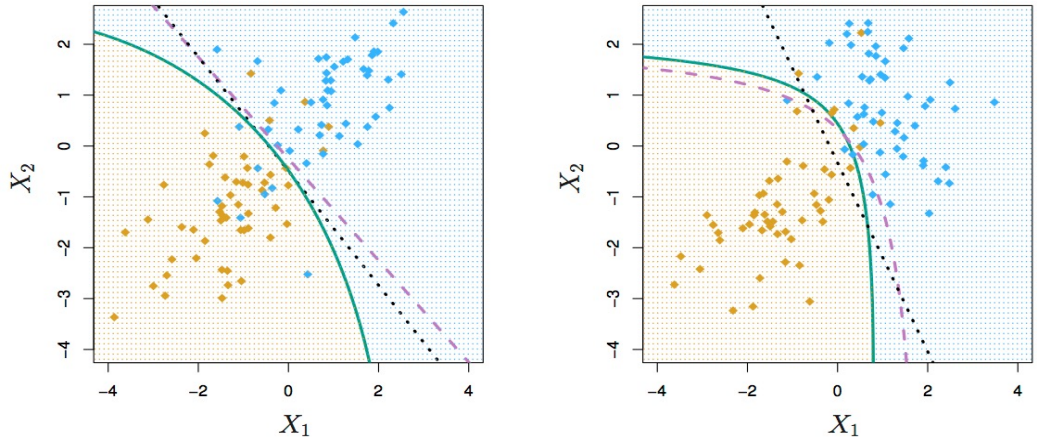
$$\vec{w} \cdot \vec{x} > c$$

Source: https://www.projectrhea.org/rhea/index.php/Discriminant_Functions_For_The_Normal(Gaussian)_Density_-_Part_2

It is often useful to see this conclusion in geometrical terms: the criterion of an input $\vec{x}$ being in a class y is purely a function of projection of multidimensional-space point $\vec{x}$ onto vector $\vec{w}$ (thus, we only consider its direction). In other words, the observation belongs to y if corresponding $\vec{x}$ is located on a certain side of a hyperplane perpendicular to $\vec{w}$. The location of the plane is defined by the threshold c.

Quadratic Discriminant Analysis (QDA)

- LDA assumed that every class has the same variance/ covariance
- QDA works identically as LDA except that it estimates separate variances/ covariance for each class
- Black dotted: LDA boundary, Green solid: QDA boundary
- Left: variances of the classes are equal (LDA is better fit)
- Right: variances of the classes are not equal (QDA is better fit)



STEVENS
INSTITUTE of TECHNOLOGY
School of Business

Thank you!

Prof. Feng Mai
School of Business

For academic use only.