PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY *of* WASHINGTON

Introduction to Machine Learning

MLEARN 510A – Lesson 3

Recap of Lesson 2

- Principle of Maximum Likelihood Estimation
- Simple Linear Regression
- Multiple Linear Regression
- Important Questions Related to a Regression Fit
- Handling Qualitative Predictors
- Linear Regression Diagnostics

Outline for Lesson 3

- Introduction to Classification
- Logistic Regression and Maximum Likelihood
- Logistic Regression Extensions
- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)
- Comparison of Various Algorithms

Classification Problems in ML

Categorize data into one of K classes

- Classify handwritten digits
- Diagnose a medical condition (stroke, drug overdose, epileptic seizure, etc.)
- Predict if transaction is fraudulent
- Determine whether given DNA mutations are disease-causing

Overview of Classification

- Input data: feature vector X and a qualitative response Y taking values in the set C
- Task: build a function f(X) that takes as input the feature vector X and predicts value for Y
- Classification algorithms estimate the probabilities that X belongs to each category in C

Why Not Linear Regression?

- Coding on categorical variable implies a natural ordering of the response variable Y

$$Y = \begin{cases} 1 & \text{if } \texttt{epileptic seizure}; \\ 2 & \text{if } \texttt{stroke}; \\ 3 & \text{if } \texttt{drug overdose}. \end{cases}$$

- Difference between drug overdose and stroke is equal to epileptic seizures?

- Average of epileptic seizure and drug overdose is equal to stroke?

Why Not Linear Regression?

- Coding scheme is not unique!

$$Y = \begin{cases} 1 & \text{if } \texttt{stroke}; \\ 2 & \text{if } \texttt{drug overdose}; \\ 3 & \text{if } \texttt{epileptic seizure}. \end{cases}$$

- Each of these coding schemes will produce a different linear regression fit

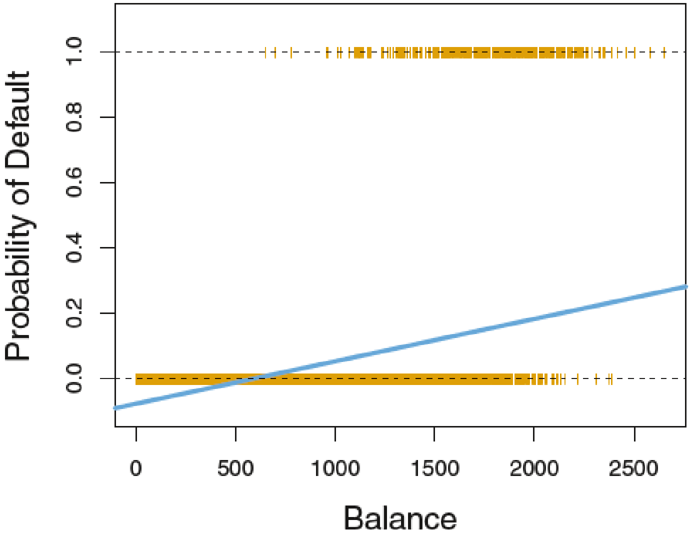
- Predictions on test set will be different too

Why Not Linear Regression? – Binary Response Variable

- Use dummy variable approach used in previous lesson

$$Y = \begin{cases} 0 & \text{if } \texttt{stroke}; \\ 1 & \text{if } \texttt{drug overdose}. \end{cases}$$

- Fit a linear regression model to this binary response
  - Predict drug overdose if Y > 0.5
  - Predict stroke otherwise



- However, estimated Y might lie outside the interval [0, 1]

- What do we mean by negative probabilities?

Quiz

- Clearly, our problem is that modeling $P(X) = P(Y = 1 \mid X) = \beta_0 + \beta_1 X$ does not work.

- What kind of function could we use to model $P(X) = P(Y = 1 \mid X)$ so that estimates of probability stay within the [0, 1] interval?

The Logistic Function

- We use a **Logistic Function**

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}.$$

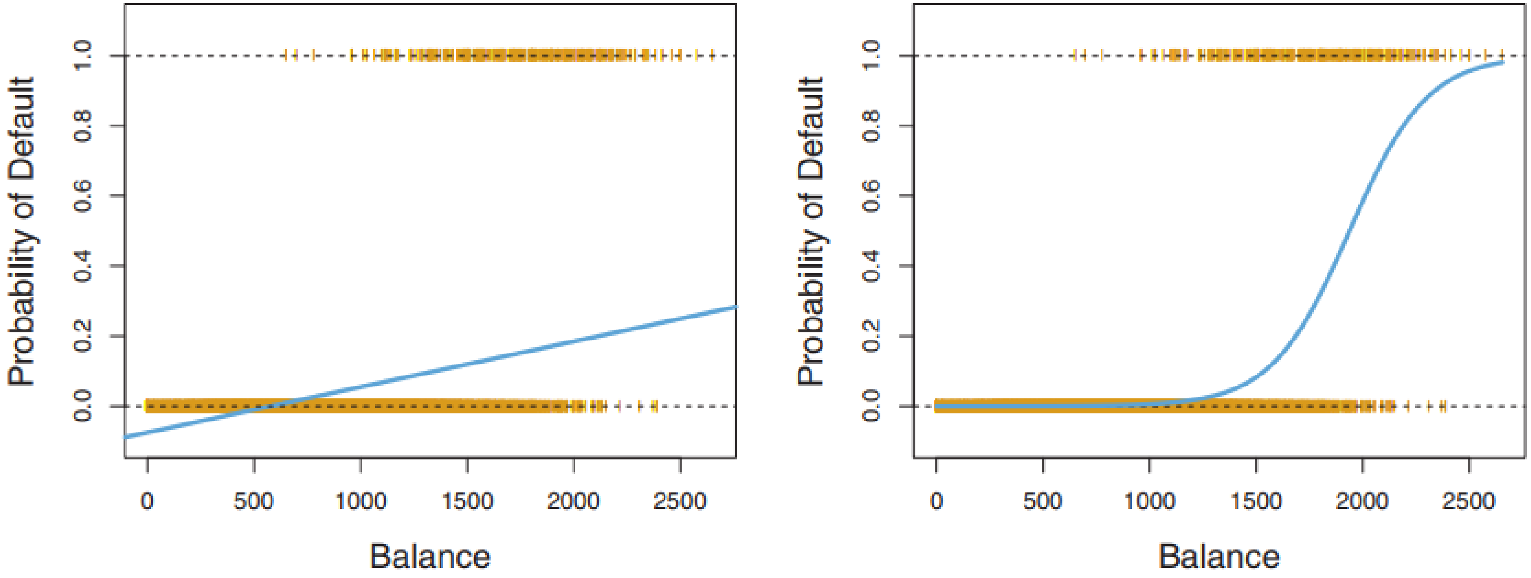
- Let’s say X is the balance on credit card and Y indicates a default

- For low balances we now predict the probability of default as close to, but never below, zero

- For high balances we predict a default probability close to, but never above, one

Linear vs. Logistic Function

$$\Pr(\texttt{default} = \texttt{Yes}|\texttt{balance})$$



Logistic Regression

- The logistic function can be expressed as

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}.$$

- This quantity is called *odds.* Small and high values indicate low and high probability of default, respectively

- Taking log on both sides, we get the *log-odds*

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X.$$

- Notice that log-odds is linear in X!

Estimating Logistic Regression Coefficients

- Coefficients are estimated using the principle of Maximum Likelihood estimation

- *We choose the parameter that maximizes the likelihood of having the obtained data at hand. With discrete distributions, the likelihood is the same as the probability. We choose the parameter for the density that maximizes the probability of the data coming from it*

- MLE requires us to maximize the likelihood function L(θ) with respect to the unknown parameter θ

$$l(\theta) = \log L(\theta) = \log \prod_{i=1}^{n} f(X_i|\theta) = \sum_{i=1}^{n} \log f(X_i|\theta)$$

- Maximizing *l*(θ) with respect to θ will give us the MLE estimator

Bernoulli Random Variable

- A Bernoulli random variable takes on values 1 and 0 with probability *p* and (1-*p*), respectively

- The binary response Y can be modeled as a Bernoulli random variable, i.e., $P(Y = 1 \mid X) = p$

- The pmf of Bernoulli random variable is $P(Y = k) = p^k(1 - p)^{(1 - k)}$

- *Exercise*
  - Verify that expectation of a Bernoulli random variable is *p*
  - Verify that variance of a Bernoulli random variable is *p(1-p)*

Maximum Likelihood Estimation

- Assume that P(Y = 1|X = x) = p(x;θ), for some function p parameterized by θ

- Assume that observations are independent of each other. The (conditional) likelihood function is

$$\prod_{i=1}^{n} \Pr\left(Y = y_i | X = x_i\right) = \prod_{i=1}^{n} p(x_i;\theta)^{y_i}(1 - p(x_i;\theta)^{1-y_i})$$

- We will maximize the likelihood function to estimates parameters $\theta=[\beta_0, \beta_1]$

Maximizing Likelihood

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

$$\begin{aligned}
\ell(\beta_0, \beta) &= \sum_{i=1}^{n} y_i \log p(x_i) + (1 - y_i) \log 1 - p(x_i) \\
&= \sum_{i=1}^{n} \log 1 - p(x_i) + \sum_{i=1}^{n} y_i \log \frac{p(x_i)}{1 - p(x_i)} \\
&= \sum_{i=1}^{n} \log 1 - p(x_i) + \sum_{i=1}^{n} y_i (\beta_0 + x_i \cdot \beta) \\
&= \sum_{i=1}^{n} - \log 1 + e^{\beta_0 + x_i \cdot \beta} + \sum_{i=1}^{n} y_i (\beta_0 + x_i \cdot \beta)
\end{aligned}$$

Exercise

- Verify that

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} (y_i - p(x_i; \beta_0, \beta))\, x_{ij}$$

Maximizing Likelihood

- The likelihood function does not have a closed form solution

- Solve it numerically using optimization methods, for e.g., Newton's Method for Numerical Optimization

$$\beta^{(n+1)} = \beta^{(n)} - H^{-1}(\beta^{(n)})\nabla f(\beta^{(n)})$$

- Matrix H is the Hessian of *f,* its matrix of partial derivatives

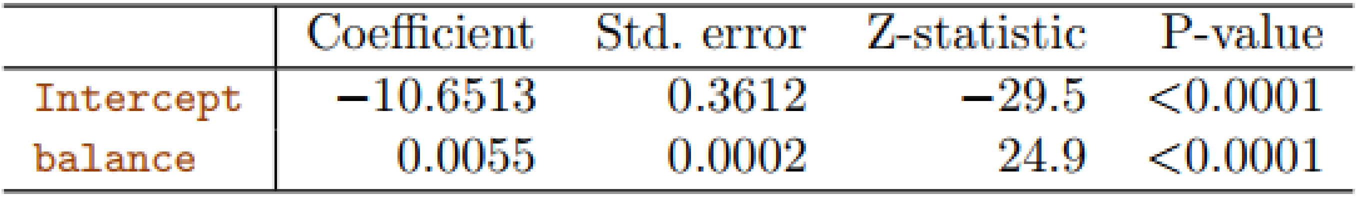
Quiz

- Can you see the bottleneck with updating β?

$$\beta^{(n+1)} = \beta^{(n)} - H^{-1}(\beta^{(n)}) \nabla f(\beta^{(n)})$$

Logistic Regression Fit

- After running numerical optimization, MLE provides an estimate of the intercept and slope parameters



	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

- The *z*-statistic in Table plays the same role as the *t*-statistic

- A large (absolute) value of the *z*-statistic indicates evidence against the null hypothesis $H_0 : \beta_1 = 0$

Interpreting Coefficients

- Linear Regression → β_1 gives the average change in Y associated with a one-unit increase in X

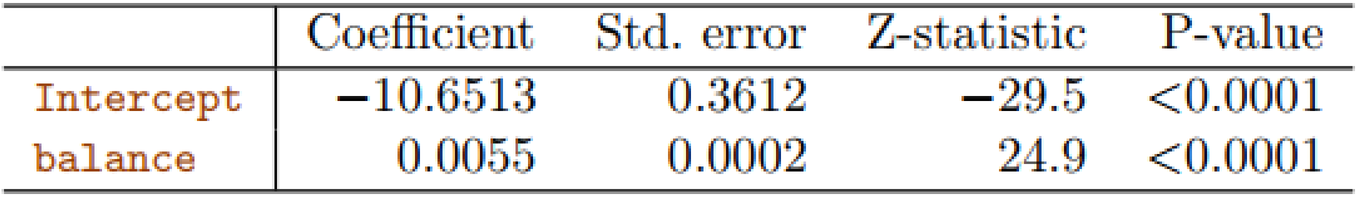
- Logistic Regression → increasing X by one unit changes the log odds by β_1, or equivalently it multiplies the odds by $e^{\beta 1}$

- If β_1 is positive, increasing X is associated with increasing $p(X)$

- If β_1 is negative, increasing X is associated with decreasing $p(X)$

Making Predictions

- Simply plug in the values to calculate probability



	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

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1,000}}{1+e^{-10.6513+0.0055\times 1,000}} = 0.00576,$$

- Probability of default for an individual with a balance of $1000 is less than 1%

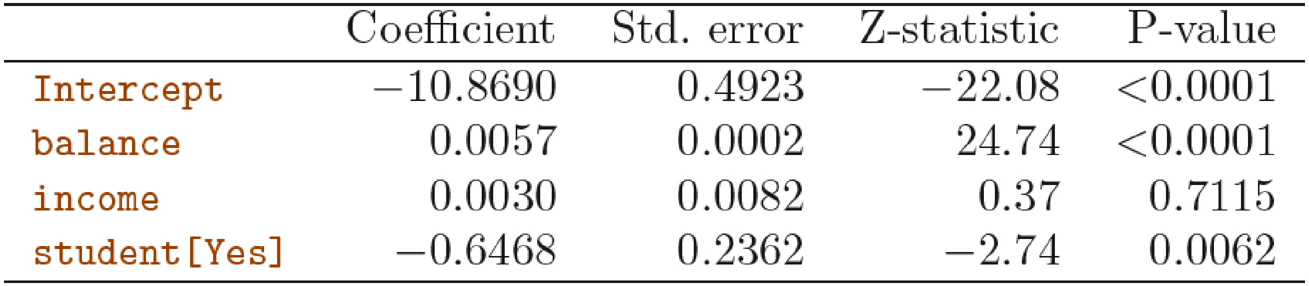
Multiple Logistic Regression

- Simply extend log-odds from simple to multiple regression

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,$$

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}.$$

- Use MLE to estimate the coefficients



	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

Multiclass Classification

- In many cases, response takes on more than two values

- Instead of having one set of parameters β_0,β_1, each class c in 0 : (k −1) will have its own offset β_0(c) and vector β_1(c)

$$\Pr\left(Y=c|\vec{X}=x\right)=\frac{e^{\beta_0^{(c)}+x\cdot\beta^{(c)}}}{\sum_c e^{\beta_0^{(c)}+x\cdot\beta^{(c)}}}$$

Generative vs. Discriminative Models

- **Generative Models** learn a model of the joint probability P(X, Y), of the inputs X and labels Y
  - Predictions are made using Bayes Rule to calculate p(Y | X)
  - Pick the most likely label Y
- **Discriminative Models** directly model p(Y | X), or learn a direct map from input X to class labels Y
  - Example: Logistic Regression
- *“One should solve the (classification) problem directly and never solve a more general problem as an intermediate step (such as modeling p (X | Y)”*

  - Vapnik

Linear Discriminant Analysis (LDA)

- Alternative approach to classification based on the theory of Generative Modeling

- Model the distribution of the predictors X separately in each of the response classes (i.e., given Y)

- Then use Bayes' theorem to flip these around into estimates for $\Pr(Y = k|X = x)$

- Similar to Logistic Regression when these distributions are assumed to be normal

Bayes Rule

- Conditional probability $\mathbb{P}(\mathrm{A|B}) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$ can be expressed as

$\mathbb{P}(A \cap B) = \mathbb{P}(\mathrm{A|B}) \times \mathbb{P}(B)$

- Conditional probability $\mathbb{P}(\mathrm{B|A}) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)}$ can be expressed as

$\mathbb{P}(B \cap A) = \mathbb{P}(\mathrm{B|A}) \times \mathbb{P}(A)$

- This implies $\mathbb{P}(\mathrm{A|B}) \times \mathbb{P}(B) = \mathbb{P}(\mathrm{B|A}) \times \mathbb{P}(A)$
- Dividing both side by $\mathbb{P}(B)$ we obtain

Bayes Rule

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

Quiz – Recalling Bayes Rule

- *You are about to get on a plane to Seattle. You want to know if you should bring an umbrella. You call 3 random friends of yours who live there and ask each independently if it's raining. Each of your friends has a 2/3 chance of telling you the truth and a 1/3 chance of messing with you by lying. All 3 friends tell you that "Yes" it is raining. What is the probability that it's actually raining in Seattle*

Why Do We Need LDA?

- Estimates for the regression coefficients are “surprisingly” unstable when the classes are well separated

- If ‘n’ is small and the distribution of predictors is approximately Gaussian, the linear discriminant model is more stable

- Linear Discriminant Analysis (LDA) is popular when we have more than two classes

Bayes Theorem for Classification

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

Where: $$f_k(x)\equiv\Pr(X=x|Y=k)$$

π_k - Prior probability

- ➢ Instead of directly computing $p_k(X)$, simply plug in estimates of π_k and $f_k(X)$
- ➢ Estimating $f_k(X)$ tends to be more challenging, unless we assume some simple forms for these densities.

LDA for p = 1

- Assume p = 1

- Estimate $f_k(X)$ by assuming a functional form

- Classify the observation to the class for which $p_k(X)$ is the greatest

- When $f_k(X)$ is Gaussian

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

LDA for p = 1 – Continued

- Assume that the variance of X is the same in all classes

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

- Assign X to the class for which $p_k(X)$ is the greatest

Exercise

- Show that finding the largest

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

is equivalent to assigning X to the class for which

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

is the largest.

The Discriminant

- The quantity

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

is called *Discriminant*

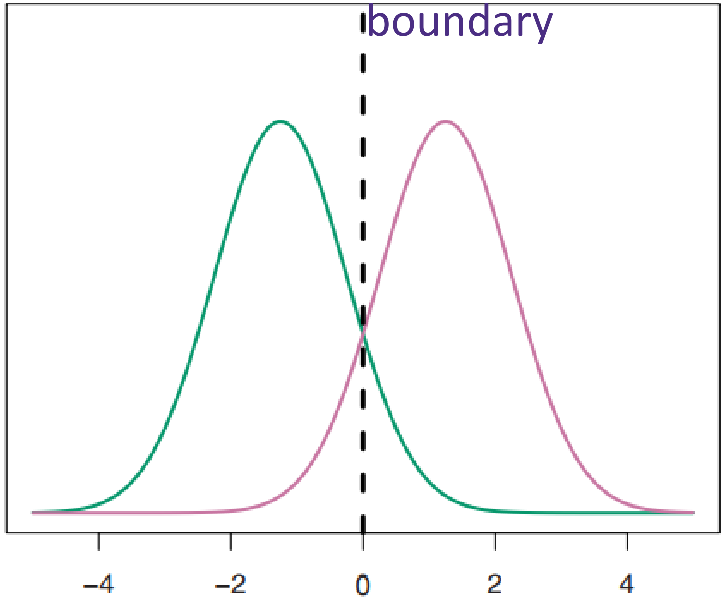
- It is a linear function of X, hence the name LDA!

Binary Class Example

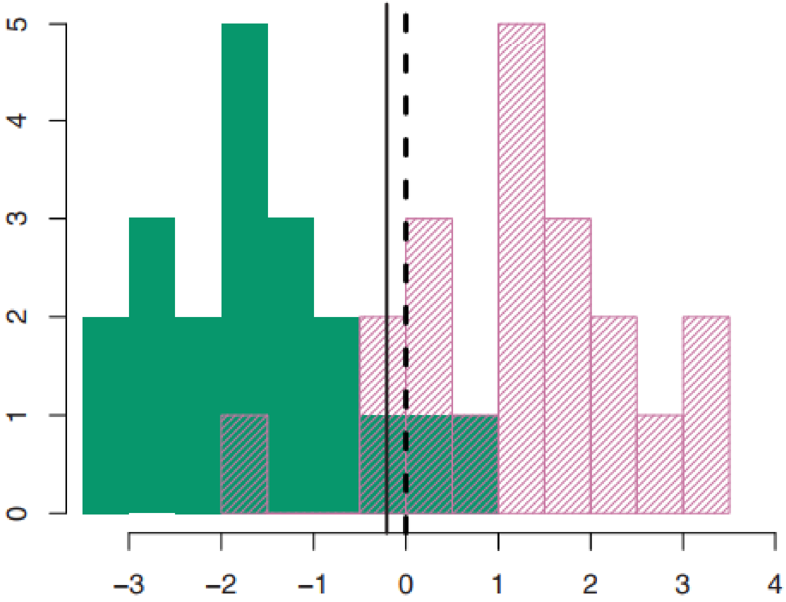
- For K = 2 and equal priors

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

$\mu_1 = -1.25$, $\mu_2 = 1.25$ $\sigma^2_1 = \sigma^2_2 = 1$



Bayes decision boundary



Unknown Means and Variances

- Sometimes, we might need to estimate the parameters $\mu_1, \ldots, \mu_K$, $\pi_1, \ldots, \pi_K$, and σ^2

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

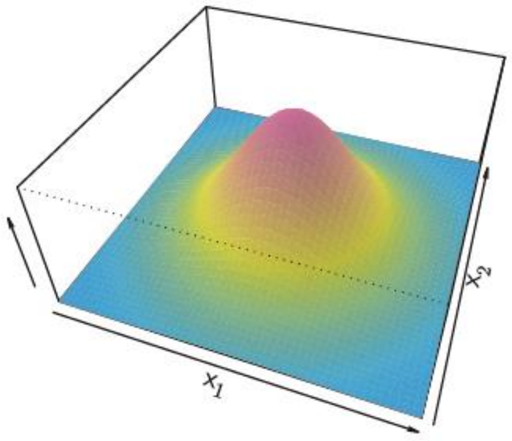
- Assign X to the class for which the estimated discriminant is the largest

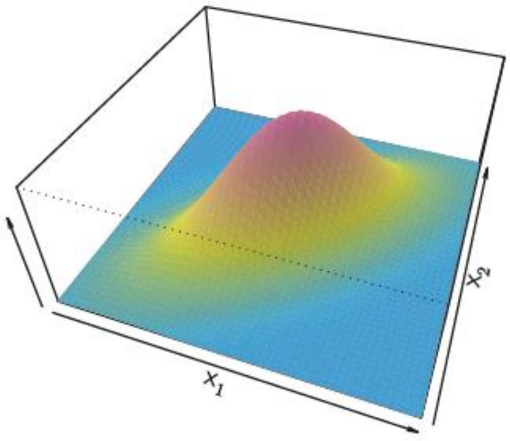
LDA for p > 1

- Assume that X = (X_1,X_2, . . .,X_p) is drawn from a multivariate Gaussian distribution, with a class-specific multivariate mean vector and a common covariance matrix

Quiz

- What can you say about the multivariate Gaussian distributions? Which plot shows random variables with equal variance?





LDA for p > 1 – Continued

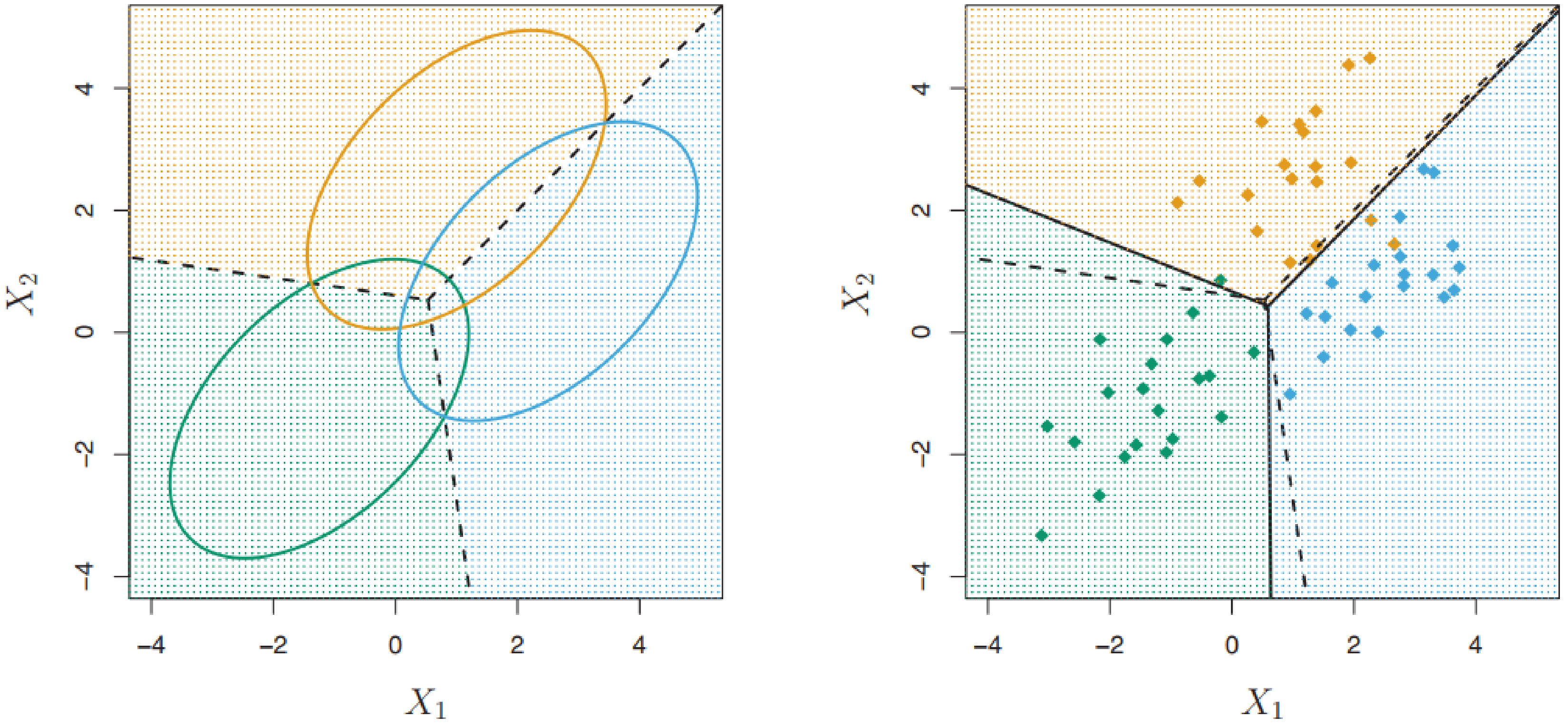
- The discriminant function in case of multivariate Gaussian distribution is given by

$$f(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\mathbf{\Sigma}^{-1}(x-\mu)\right)$$

$$\delta_k(x) = x^T\mathbf{\Sigma}^{-1}\mu_k - \frac{1}{2}\mu_k^T\mathbf{\Sigma}^{-1}\mu_k + \log\pi_k$$

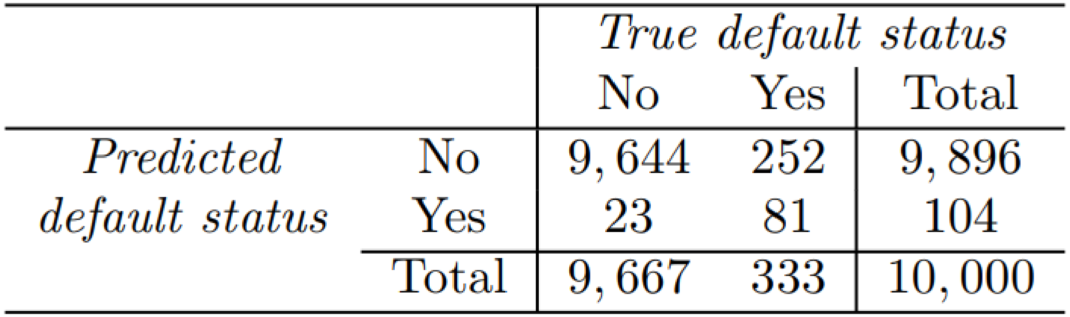
- Is this still linear?

LDA for p > 1 – Continued



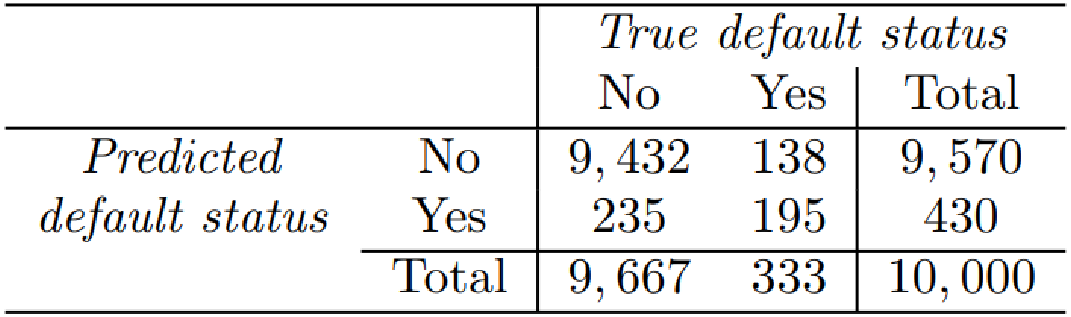
Confusion Matrix for LDA

Classification threshold
P(Default = Yes | X = x) > 0.5



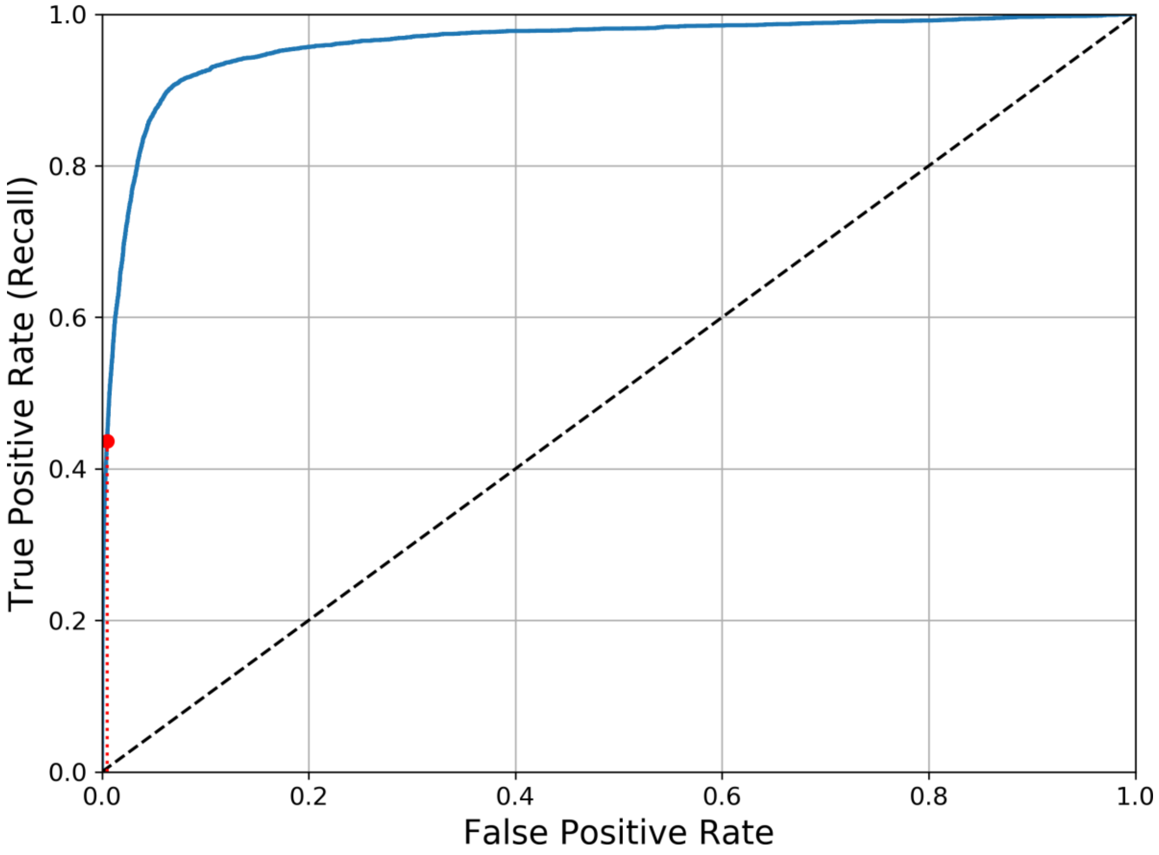
		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

Classification threshold
P(Default = Yes | X = x) > 0.2



		True default status		
		No	Yes	Total
Predicted default status	No	9,432	138	9,570
	Yes	235	195	430
	Total	9,667	333	10,000

ROC Plot


1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate (Recall)
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate

Quadratic Discriminant Analysis (QDA)

- **LDA**: Assume that X = (X_1,X_2, . . .,X_p) is drawn from a multivariate Gaussian distribution, with a class-specific multivariate mean vector and a common covariance matrix

- **QDA**: Assume that X = (X_1,X_2, . . .,X_p) is drawn from a multivariate Gaussian distribution, with a class-specific multivariate mean vector and a *class-specific* covariance matrix

Discriminant Function for QDA

- Assumes that an observation from the k^{th} class is of the form $X \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

- Bayes classifier assigns an observation $X = x$ to the class for which discriminant is the largest

$$\begin{aligned}\delta_k(x) &= -\frac{1}{2}(x-\mu_k)^T \boldsymbol{\Sigma}_k^{-1}(x-\mu_k) - \frac{1}{2}\log|\boldsymbol{\Sigma}_k| + \log \pi_k \\ &= -\frac{1}{2}x^T \boldsymbol{\Sigma}_k^{-1} x + x^T \boldsymbol{\Sigma}_k^{-1}\mu_k - \frac{1}{2}\mu_k^T \boldsymbol{\Sigma}_k^{-1}\mu_k - \frac{1}{2}\log|\boldsymbol{\Sigma}_k| + \log \pi_k\end{aligned}$$

Quiz

- Why is it called QDA?

$$\begin{aligned}\delta_k(x) &= -\frac{1}{2}(x-\mu_k)^T\boldsymbol{\Sigma}_k^{-1}(x-\mu_k) - \frac{1}{2}\log|\boldsymbol{\Sigma}_k| + \log\pi_k \\ &= -\frac{1}{2}x^T\boldsymbol{\Sigma}_k^{-1}x + x^T\boldsymbol{\Sigma}_k^{-1}\mu_k - \frac{1}{2}\mu_k^T\boldsymbol{\Sigma}_k^{-1}\mu_k - \frac{1}{2}\log|\boldsymbol{\Sigma}_k| + \log\pi_k\end{aligned}$$

LDA vs. QDA

- For *p* predictors, the LDA model estimates *Kp* linear coefficients to estimate while QDA estimates *Kp(p+1)/2* coefficients

- LDA is a much less flexible classifier than QDA → low variance

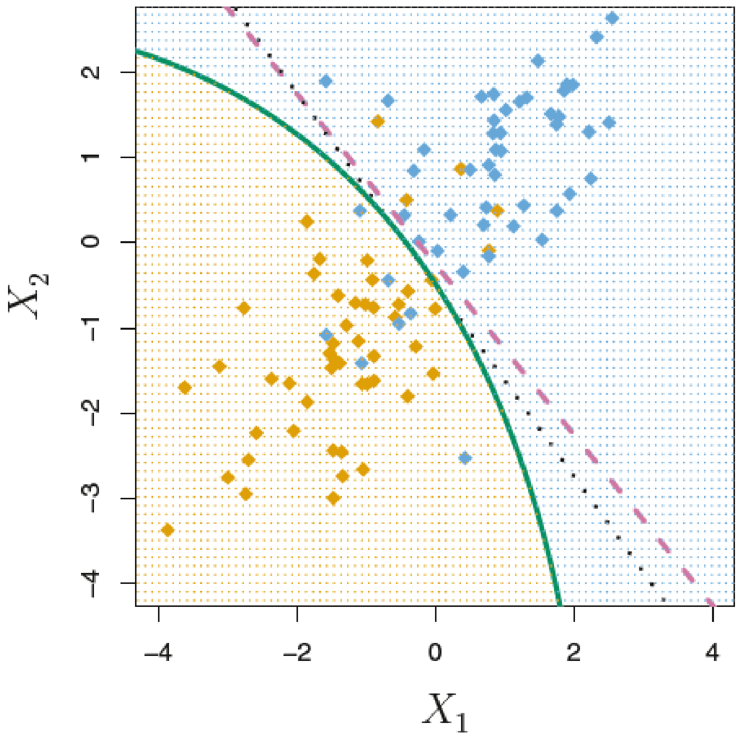
- Trade-off: if LDA's assumption that the *K* classes share a common covariance matrix is off, LDA can suffer from high bias

- Few training examples → Use LDA to reduce variance

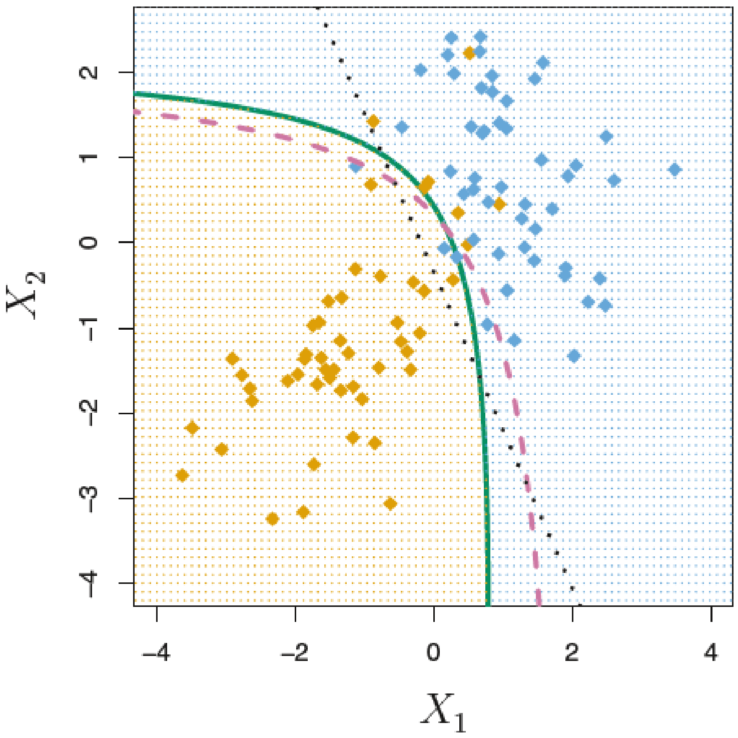
- Large training set → Use QDA since the assumption of a common covariance matrix for the *K* classes is unrealistic

LDA vs. QDA

LDA assumptions valid



LDA assumptions invalid



Purple – Bayes decision boundary

Black dotted – LDA

Green – QDA

Logistic Regression vs. LDA

- Both produce linear decision boundaries
- Logistic Regression estimates parameters using MLE
- LDA estimates parameters of the boundary using estimated means and covariances from a normal distribution
- Logistic Regression outperforms LDA when LDA assumptions do not hold

QDA vs. Logistic Regression and KNN

- QDA serves as a compromise between the non-parametric KNN method and the linear LDA and logistic regression approaches
- QDA can accurately model a wider range of problems than linear methods
- Though not as flexible as KNN, QDA can perform better in the presence of a limited number of training observations
- *No Free Lunch – There is no one model that works best for every problem!*

Jupyter Notebook

- *Case Study*

ON-BRAND STATEMENT

FOR GENERAL USE

> What defines the students and faculty of the University of Washington? Above all, it’s our belief in possibility and our unshakable optimism. It’s a connection to others, both near and far. It’s a hunger that pushes us to tackle challenges and pursue progress. It’s the conviction that together we can create a world of good. And it’s our determination to Be Boundless. Join the journey at **uw.edu**.