NYC DATA SCIENCE ACADEMY

Python Machine Learning Class 2: Classification Part I

Data Science Bootcamp

Outline

- **Limitation of Linear Regression**
  - **Logistic Regression**
  - **Discriminant Analysis: Motivation**
  - **Discriminant Analysis: Models**
  - **Naive Bayes**

Classification Problems

- *Categorical* (qualitative) variables: takes values in a finite set (usually unordered).
  - email: {spam, non-spam}
  - blood type: {A, B, AB, O}
  - tumor: {malignant, benign}
- *Classification*: given a feature (or a set of features), we want to predict categorical a output.
- Sometime people are also interested in estimating the probabilities that *an observation* belongs to each category.

A Classification Example

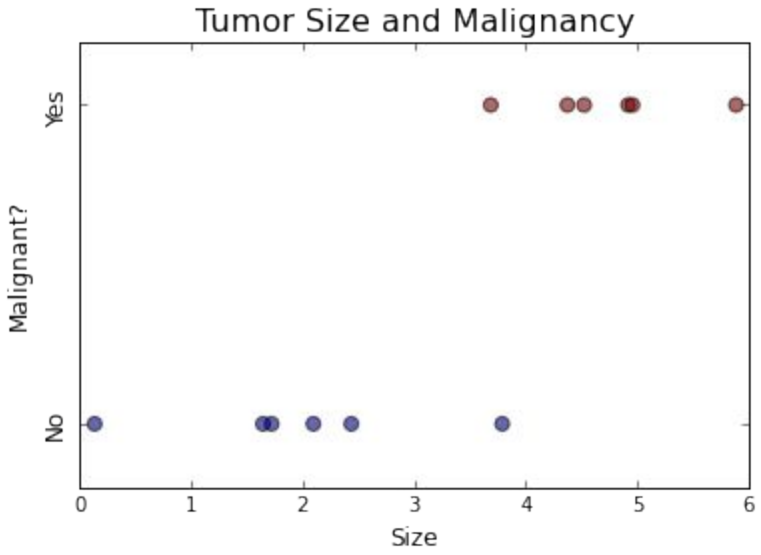
- Predict whether a tumor is malignant or benign based on the size
- The output is binary:
  - 0: benign
  - 1: malignant
- Here is a simulated data set:

	Size	Malignant
0	3.788628	0
1	2.436510	0
2	2.096497	0
3	0.136507	0
4	1.722612	0
5	1.645241	0

	Size	Malignant
6	4.917259	1
7	4.372999	1
8	4.956182	1
9	4.522782	1
10	3.686135	1
11	5.884622	1

A Classification Example

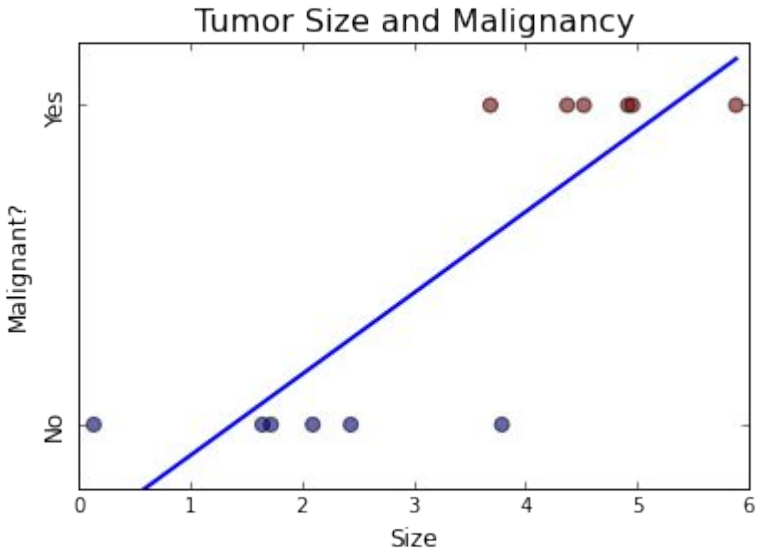
- By selecting the Size as feature and Malignant as output, we can visualize the data as:



- Question: can we use linear regression?

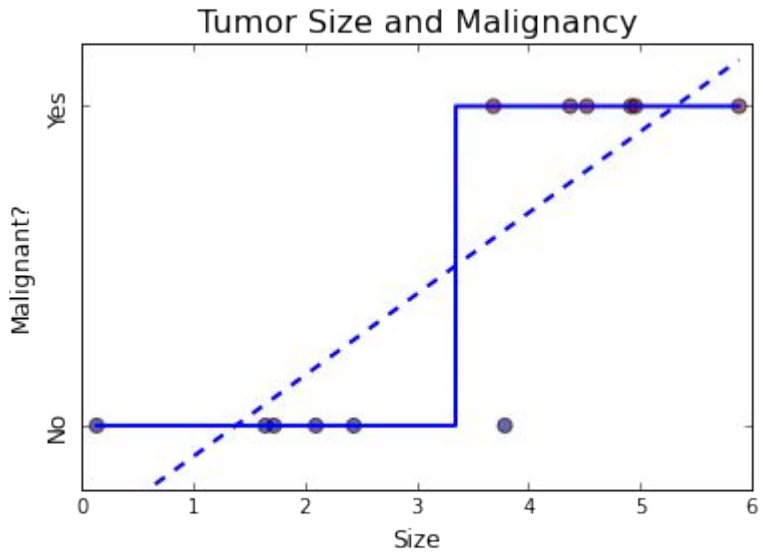
Can We Use Linear Regression?

- Let's fit a linear regression model with the simulated tumor data:



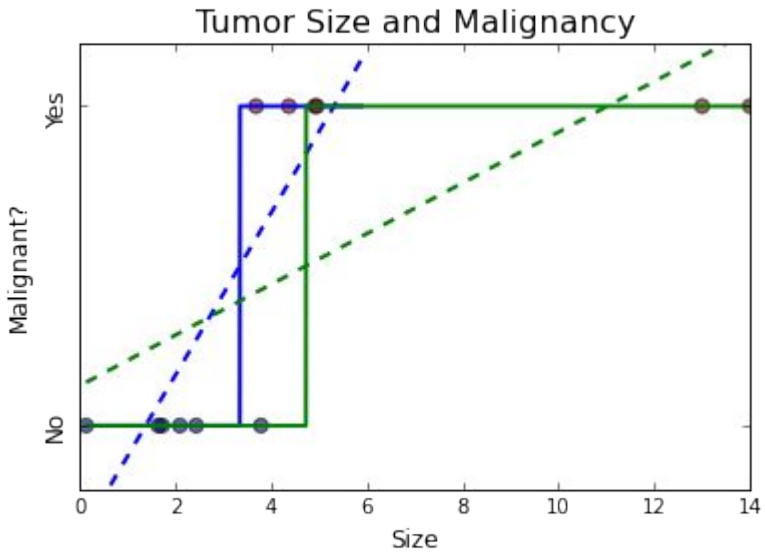
Can We Use Linear Regression?

- We may then set a threshold
  - Predict 1 if $\hat{y} \geq 0.5$
  - Predict 0 if $\hat{y} < 0.5$
- The predicted values become binary:



Issues with Linear Regression

- It looks like the binary prediction with linear regression is not so bad. However, we do have the following two problems:
  - the continuous output exceeds the interval [0, 1]. Therefore we cannot interpret it as probability.
  - the prediction can be affected by outliers easily.



Outline

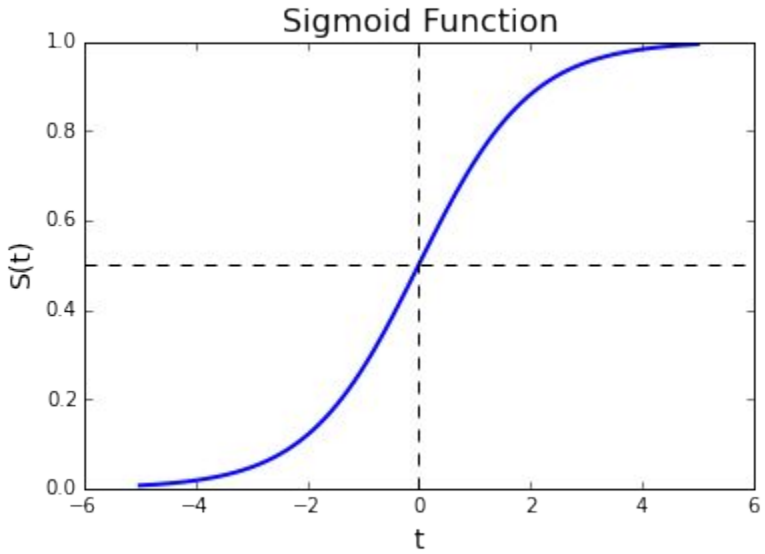
- Limitation of Linear Regression
- **Logistic Regression**
- Discriminant Analysis: Motivation
- Discriminant Analysis: Models
- Naive Bayes

Sigmoid Function

- Sigmoid Function: a monotonically increasing function which maps a real value to a new value bounded between 0 and 1.

$$S(t) = \frac{e^t}{1 + e^t}$$

- ***e*** = 2.718 is a mathematical constant (Euler's number).



Logistic Regression

Logistic regression, despite its name, is a linear model for classification rather than regression.

- Idea: if we transform the linear function $\beta_0+\beta_1 X$ using the sigmoid function $S(t)$, then no matter what values β_0, β_1 or X take, y will always have values between 0 and 1.
- *Logistic Regression models* use this form to estimate the probability that $y = 1$ given its size X:

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

- So different β_0 and β_1 will give different estimations Pr.

Maximum Likelihood

- Let's write the likelihood function

$$p(x_i, \beta_0, \beta_1) = Pr(Y = 1|X = x_i)$$

to describe the probability of observed outcomes to be 1 given $X = x_i$.

- Then, given an input X with n observations, the likelihood gives the probability of having the same label for each observations:

$$L(\beta_0, \beta_1) = \prod_{i,y_i=1} p(x_i, \beta_0, \beta_1) \prod_{i,y_i=0} (1-p(x_i, \beta_0, \beta_1))$$

where the first product gives the probability of successfully predicting the "1"s and the second product is the probability of successfully predicting the "0"s in the training data.

Maximum Likelihood

- The likelihood function $L(\beta_0, \beta_1)$ gives the probability of making the same prediction as the observed data.
- We want to pick β_0 and β_1 to maximize the likelihood $L(\beta_0, \beta_1)$, i.e., to maximize the "agreement" of the selected model with the observed data.

Log-Likelihood

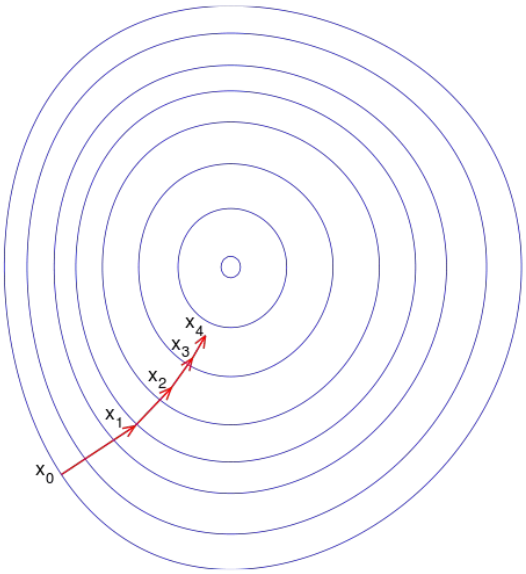
- In practice it is often more convenient to work with the logarithm of the likelihood function, called the **log-likelihood**:

$$\log L(\beta_0, \beta_1) = \sum_{i=1}^{n} \{y_i \log p(x_i, \beta_0, \beta_1) + (1 - y_i) \log(1 - p(x_i, \beta_0, \beta_1))\}$$
$$= \sum_{i=1}^{n} \{y_i(\beta_0 + \beta_1 X) - \log(1 + e^{\beta_0 + \beta_1 X})\}$$

- Logistic regression models are usually fit by maximum likelihood, i.e., to finds β_0 and β_1 that maximize the function above.

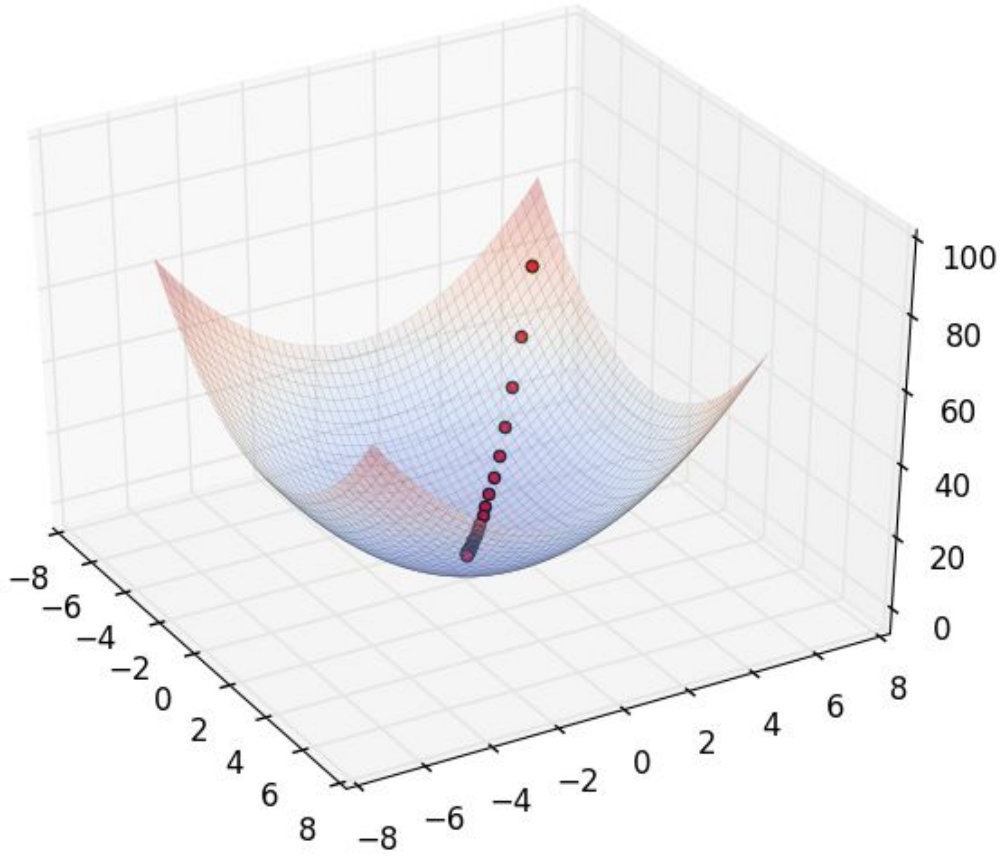
Gradient Descent

- To maximize the log-likelihood, most packages, including scikit-learn, use a numerical method called *gradient descent*, i.e., to find the maximum by search along a steepest path on the log-likelihood function.



source: https://en.wikipedia.org/wiki/Gradient_descent

Gradient Descent



Making Predictions

- After estimating the parameters, the likelihood function $p(x, \beta_0, \beta_1)$ predicts the probability of output Y to be 1 given x. If we set a threshold, then we can predict binary outputs.
- Let's consider using tumor size to predict malignancy:
  - If the maximum likelihood gives $Pr(Y=1|X=x) = 0.2$, then our prediction is 20% chance of tumor being malignant, or equivalently, 80% of change it's benign.
  - If we set the threshold, for example, to be 0.5, then we can predict the tumor is benign.

Hands-on Session

- Please go to the **"Logistic Regression in Scikit-Learn"** in the lecture code.

Outline

- Limitation of Linear Regression
- Logistic Regression
- **Discriminant Analysis: Motivation**
- Discriminant Analysis: Models
- Naive Bayes

Conditional Probability

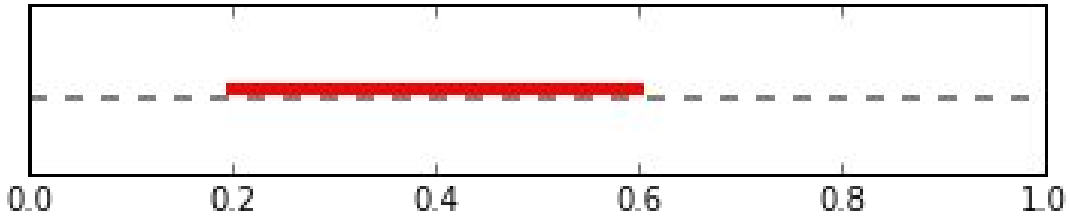
- Let *Y* be an event with probability *P*(*Y*) > 0, the *conditional probability* of observing *X* given that *Y* has occurred is defined as:

$$P(X|Y) = \frac{P(X,Y)}{P(Y)}$$

  - *P*(*X*, *Y*) refers to the probability that *X* and *Y* occur at the same time.
  - *P*(*X*|*Y*) is the probability of *X* after insuring *Y*'s occurrence.

Conditional Probability

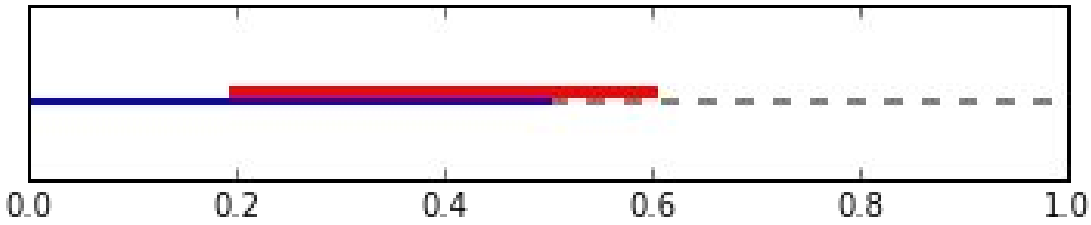
- How do we compute the probability of obtaining a number in the red region?



$$0.4 \div 1 = 0.4$$

Conditional Probability

- How about the probability of obtaining a number in the red region when restricted in the blue region?



$$\frac{P(\text{red, blue})}{P(\text{blue})} = \frac{0.3}{0.5} = 0.6$$

Conditional Probability

- If *X* and *Y* are independent, $P(X, Y) = P(X)P(Y)$:
  - Then the conditional probability is

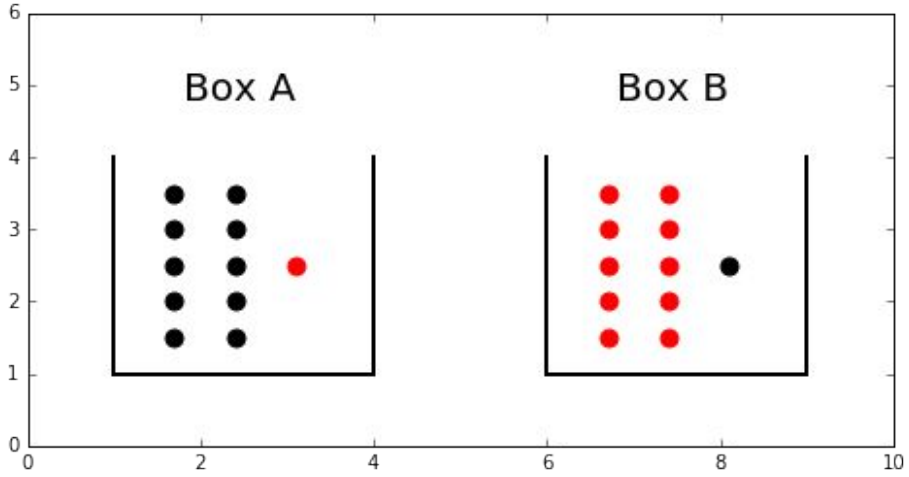
$$P(X|Y) = \frac{P(X,Y)}{P(Y)} = \frac{P(X)P(Y)}{P(Y)} = P(X)$$

  - This implies that the occurrence of *Y* does not have any impact on *X*.

Conditional Probability Example

Consider an experiment of picking balls of two colors, red and black, from two boxes labeled A and B.

1. There are 10 black balls and 1 red ball in box A, and 1 black ball and 10 red balls in box B.
2. We randomly choose a box and then pick a ball from it.
3. What is the probability that we get a red ball?



Conditional Probability Example

When choosing a box to pick, we have:

1. $P(A) = P(B) = 0.5$.
2. If we choose A, $P(red|A) = 1/11$.
3. If we choose B, $P(red|B) = 10/11$.

So the probability to get one red ball from either box A or box B is:

$$\begin{aligned} P(red) &= P(red|A) \cdot P(A) + P(red|B) \cdot P(B) \\ &= \frac{1}{11} \times 0.5 + \frac{10}{11} \times 0.5 \\ &= \frac{1}{2} \end{aligned}$$

Bayes Theorem

- Bayes theorem is named after Thomas Bayes.
- It describes the probability of an event, based on conditions that might be related to the event.
- Bayes theorem states:

$$Pr(Y|X) = \frac{Pr(X|Y) \cdot Pr(Y)}{Pr(X)}$$
$$= \frac{Pr(X|Y) \cdot Pr(Y)}{\sum_l Pr(X|Y = l) \cdot Pr(Y = l)}$$

Bayes Theorem Example

- Consider the same experiment of picking balls from two boxes.
- If the ball we picked is red, then what is the probability that the ball was from box *A*?

According to Bayes' theorem, we have:

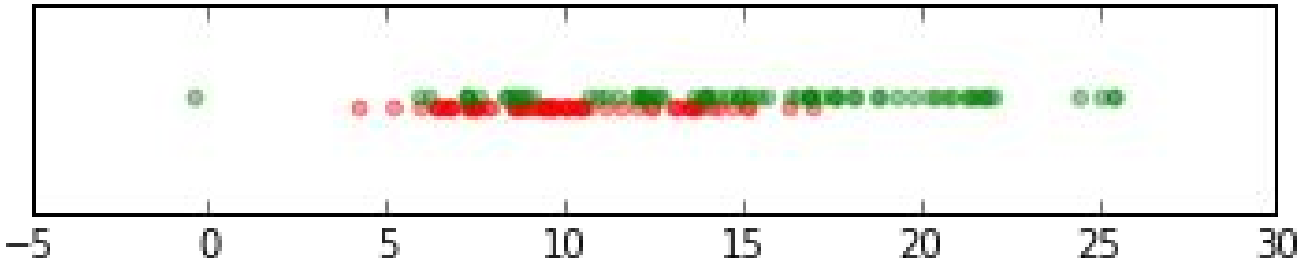
$$
\begin{aligned}
P(A|red) &= \frac{P(red|A) \cdot P(A)}{P(red)} \\
&= \frac{P(red|A) \cdot P(A)}{P(red|A)P(A) + P(red|B)P(B)} \\
&= \frac{\frac{1}{11} \times 0.5}{(\frac{1}{11} \times 0.5 + \frac{10}{11} \times 0.5)} \\
&= \frac{1}{11}
\end{aligned}
$$

Bayes Theorem Example

- How does this relate to our classification problem? Consider from the training set we realize that for a red ball:
  - the probability that the red ball was from box A is 1/11
  - and the probability that the red ball was from box B is 10/11
- Next time if we get a red ball, shouldn't we be more confident that the ball was from box B?

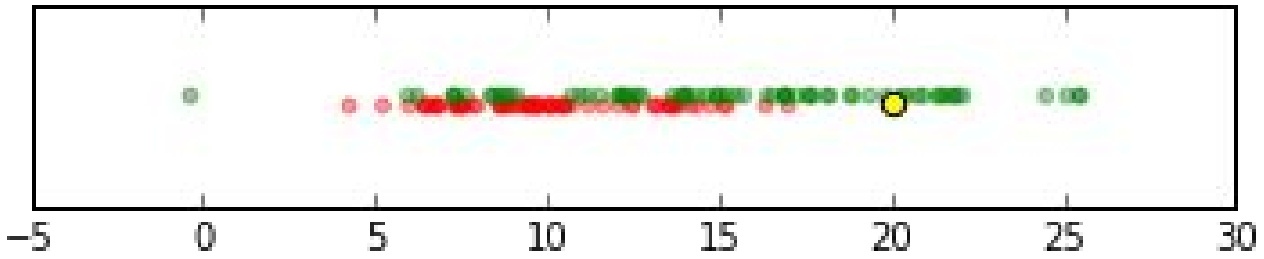
Discriminant Analysis

- *Discriminant analysis* is a statistical analysis technique which finds a set of prediction equations based on independent variables that are used to classify observations into classes.
- **Motivation**: To be more precisely, let's consider binary classification based on a numerical feature.



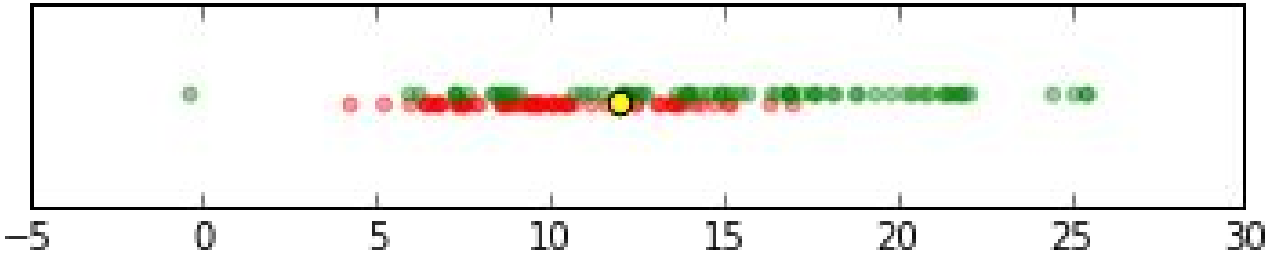
Discriminant Analysis

- If we found a new observation, which class do you think it belongs to?



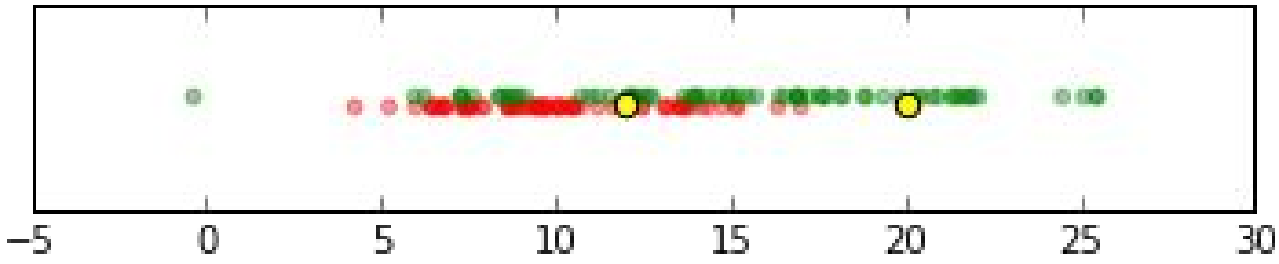
Discriminant Analysis

- What about this one?



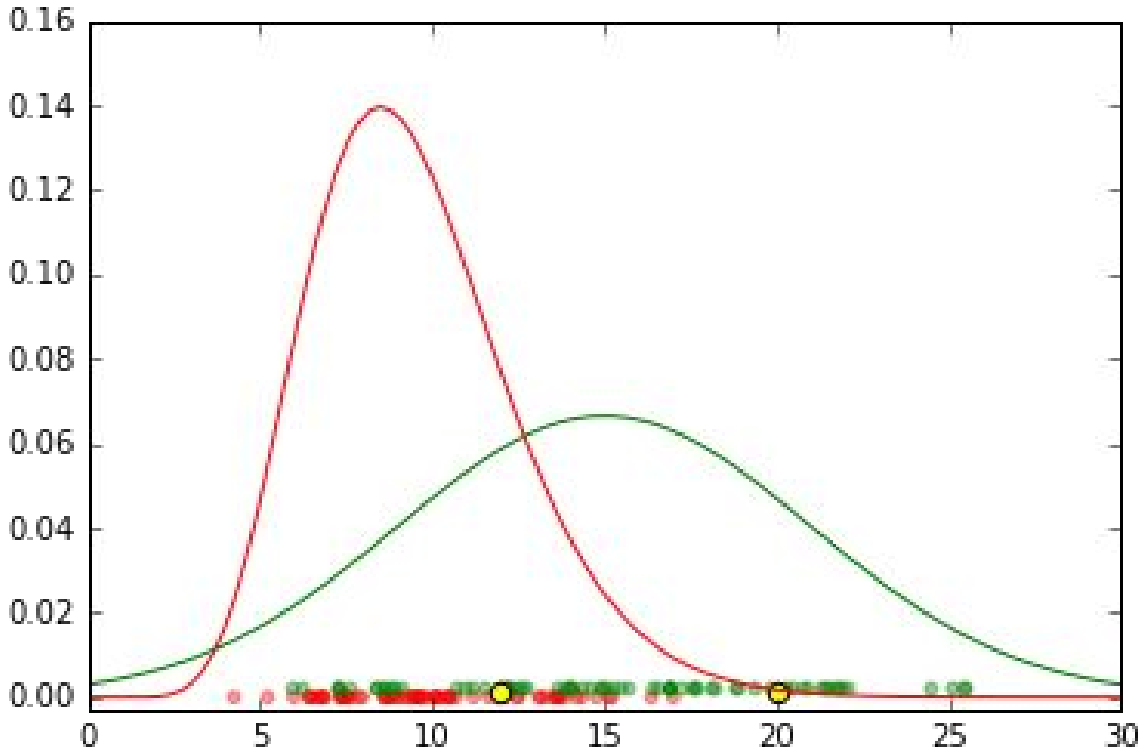
Discriminant Analysis

- What makes us feel differently?
  - If there is some other information tacitly guided us to the conclusion, can we somehow name it? Visualize it?



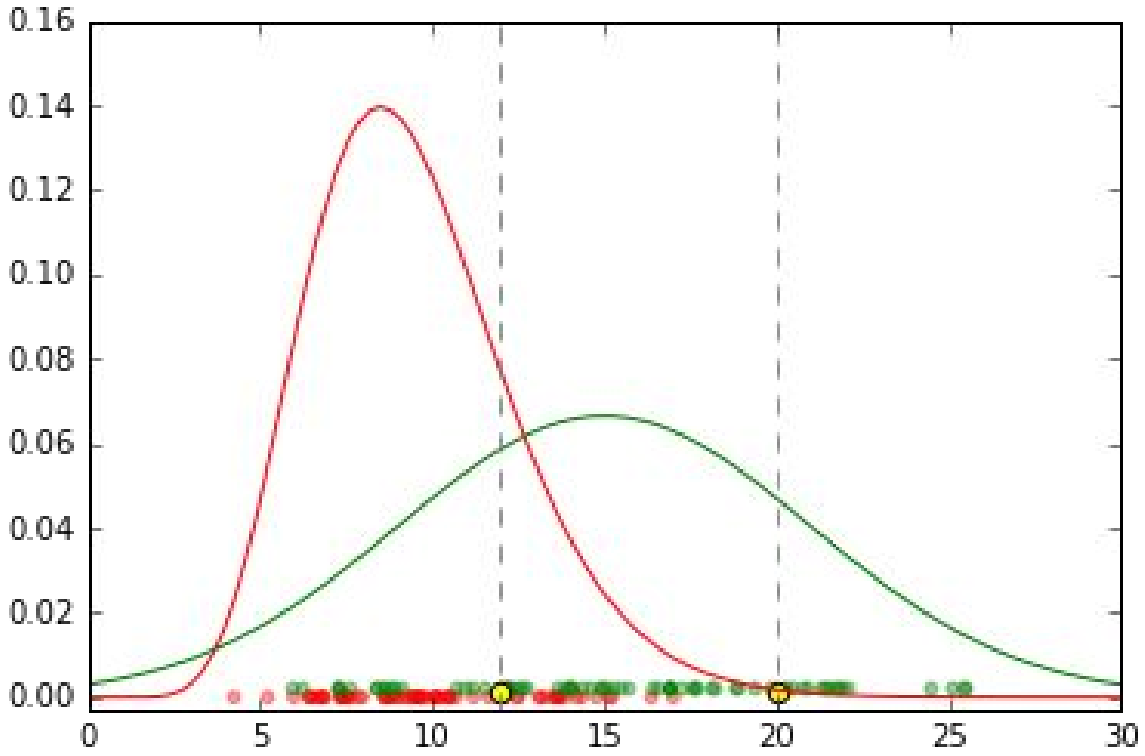
Discriminant Analysis

- How about density plot for each class?



Discriminant Analysis

- What happens to the density plots at the two yellow observations?



Discriminant Analysis

- Note the goal of classification is to compute

$$P(Y = k \mid X = x) \text{ for each class k}$$

- But we just found that

$$p(X = x \mid Y = k) \text{ for each class k}$$

Is helpful! How do we relate the two kinds of conditional probability?

Discriminant Analysis and Bayes Theorem

- Bayes theorem comes into play because we want to relate the two conditional probabilities above.

$$P(Y = k \mid X = x) = \frac{p(X = x \mid Y = k)P(Y = k)}{\sum_l p(X = x \mid Y = l)P(Y = l)}$$

- **Questions:**
  - How do we model $P(Y = k)$ (this is called the prior probability for class k)?
  - How do we model $p(X = x \mid Y = k)$?

Discriminant Analysis and Bayes Theorem

- **Answers:**

  - $P(Y = k)$ Is simply $\frac{n_k}{n}$.

  Where:

  - n_k = the number of observations in class k.
  - n = the total count of observations.

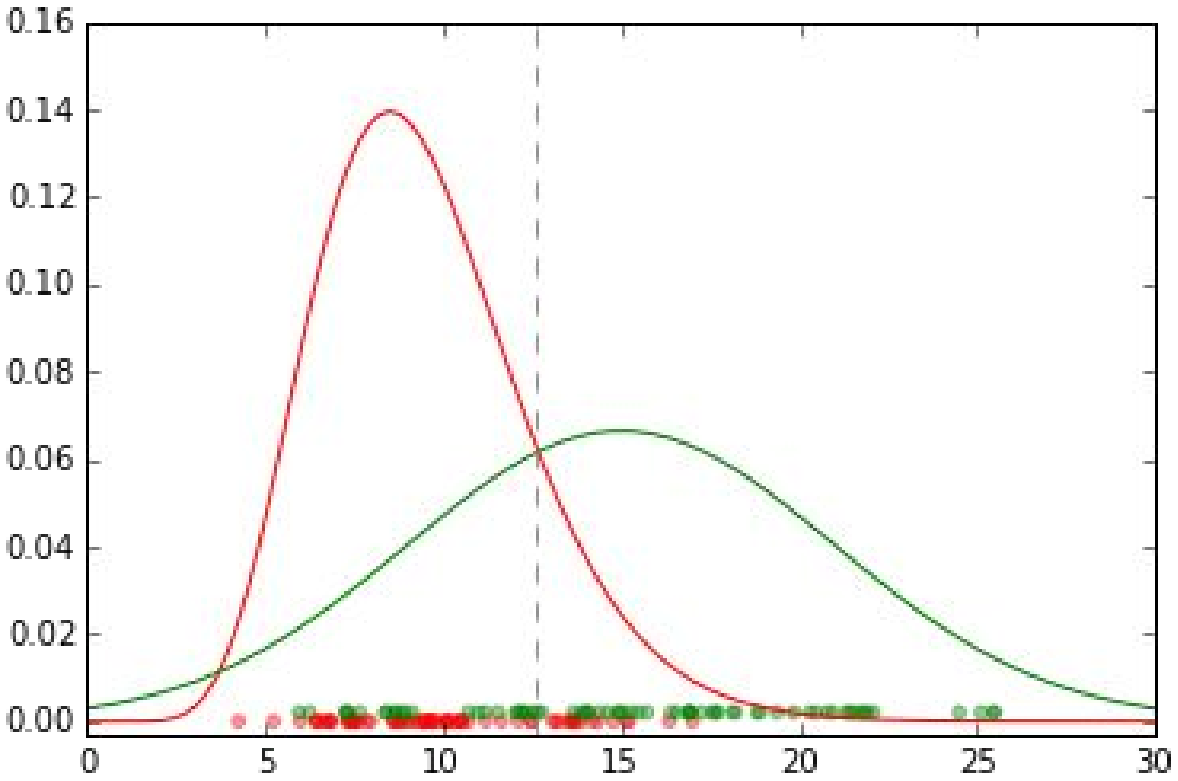
  - Modeling $p(X = x \mid Y = k)$ is nontrivial. Different models result in different classifiers as we will see.

Bayes Classifier

- Now that we can predict the probability of belonging to a particular class, we can then classify the observation to the class with the highest probability.
  - This is known as Bayes classifier. It minimises the probability of misclassification.
  - The boundary of classification is simply where the probability of different classes happen to be the same.

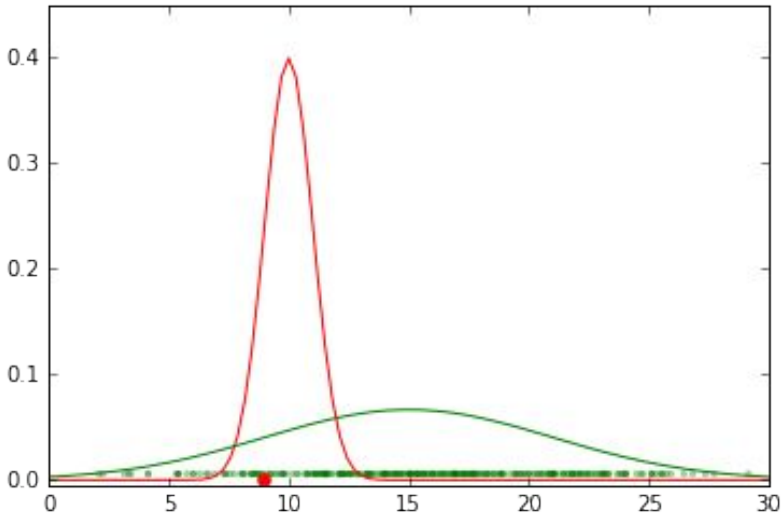
Bayes Classifier

- So is this how we classify?



Bayes Classifier

- **Observation:** There are prior probabilities in discriminant analysis, but the boundary above did NOT consider them.
  - To emphasize the effect of the density within each class, we intentionally created two classes with the same size. When the sizes are different, **missing the prior would cause a big trouble**.
  - Below is an extreme case:



Outline

- **Limitation of Linear Regression**
  - Logistic Regression
  - Discriminant Analysis: Motivation
- **Discriminant Analysis: Models**
  - Naive Bayes

Discriminant Analysis: Models

- To build a Bayes classifier, the only thing we miss is

$$p(X = x \mid Y = k)$$

- Since this is a continuous distribution, the Gaussian distribution is widely used to model it. Different kinds of Gaussian distribution result in different kind of classifiers. The following three are most common:
  - **Linear Discriminant Analysis** (LDA)
  - **Quadratic Discriminant Analysis** (QDA)
  - **Gaussian Naive Bayes** (This is the same as QDA in a one dimensional case)

Outline

- Limitation of Linear Regression
  - Logistic Regression
  - Discriminant Analysis: Motivation
- Discriminant Analysis: Models
  - One Dimensional Cases
  - Naive Bayes

One Dimensional Cases

- When we have only one feature, we use one dimensional Gaussian distribution.

$$N(\mu, \sigma)(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

  - Note that it is sufficient to specify the mean and the standard deviation to specify a Gaussian distribution.

One Dimensional Cases

- We always allow **different means** among different classes, but....

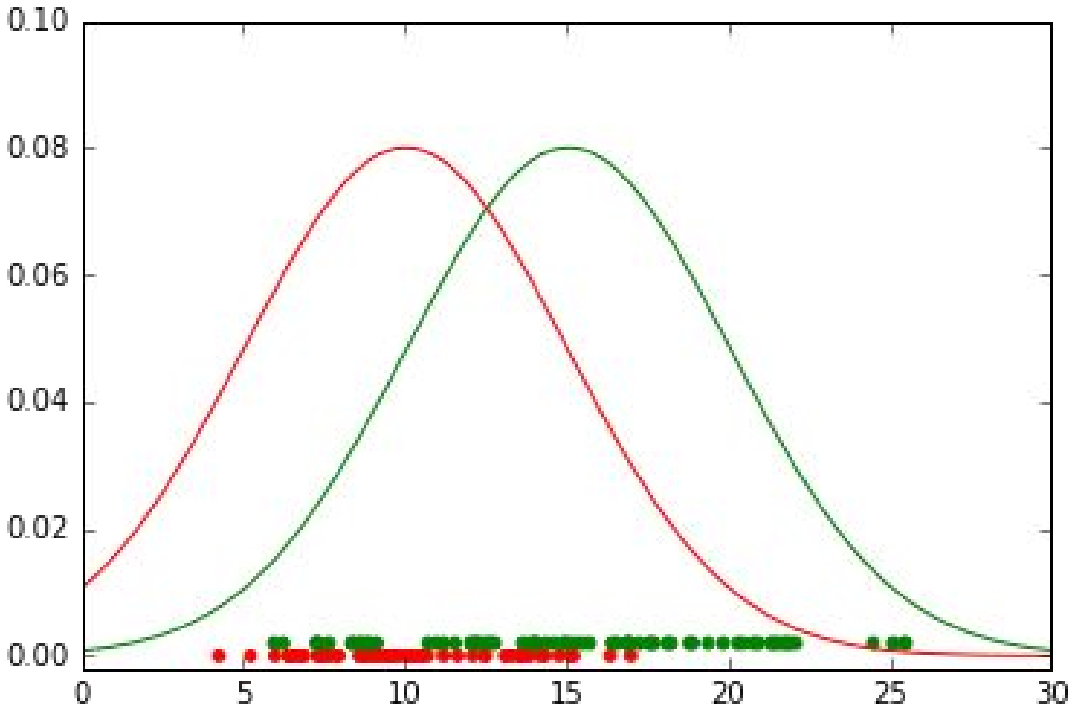
Linear Discriminant Analysis

- For LDA, we assume that the standard deviation is the same for every class. In one dimensional case, this means that the distribution for each class k is:

$$p(X = x \mid Y = k) = \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{1}{2}\left(\frac{x - \mu_k}{\sigma}\right)^2\right]$$

Linear Discriminant Analysis

- With visualization, this means the width of the distribution for every class is unchanged.



Linear Discriminant Analysis

- **Question:** Now we know that with LDA the distribution for each class k is:

$$p(X = x \mid Y = k) = \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{1}{2}\left(\frac{x - \mu_k}{\sigma}\right)^2\right]$$

  - How do we decide μ_k and σ?

Linear Discriminant Analysis

- **Answer:**

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

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

$$= \sum_{k=1}^{K} \frac{n_k - 1}{n-K} \cdot \hat{\sigma}_k^2$$

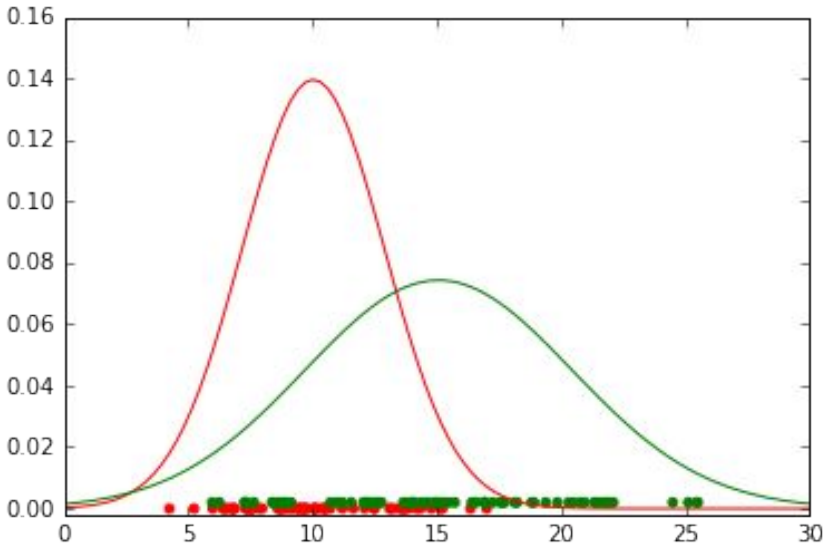
where

- K is the total number of classes.
- $\hat{\sigma}_k^2 = \frac{1}{n_k - 1} \sum_{i;y_i=k} (x_i - \hat{\mu}_k)^2$ is the sample variance of class k.

Quadratic Discriminant Analysis

- For QDA, the standard deviation can vary among the classes. In one dimensional case, this means the width of the distribution for every class can be different. Therefore:

$$p(X = x \mid Y = k) = \frac{1}{\sqrt{2\pi}\sigma_k} exp\Big[-\frac{1}{2}\Big(\frac{x - \mu_k}{\sigma_k}\Big)^2\Big]$$



Quadratic Discriminant Analysis

- **Question:** Now we know that with LDA the distribution for each class k is:

$$p(X = x \mid Y = k) = \frac{1}{\sqrt{2\pi}\sigma_k} exp\left[-\frac{1}{2}\left(\frac{x - \mu_k}{\sigma_k}\right)^2\right]$$

  - How do we estimate $\hat{\mu}_k$ and $\hat{\sigma}_k$?

Outline

- Limitation of Linear Regression
  - Logistic Regression
  - Discriminant Analysis: Motivation
  - Discriminant Analysis: Models
    - One Dimensional Cases
    - **Higher Dimensional Cases**
  - Naive Bayes

Higher Dimensional Cases

- We start with the discussion on higher dimensional Gaussian distribution. This is essentially the only difference in higher dimensional discriminant analysis.

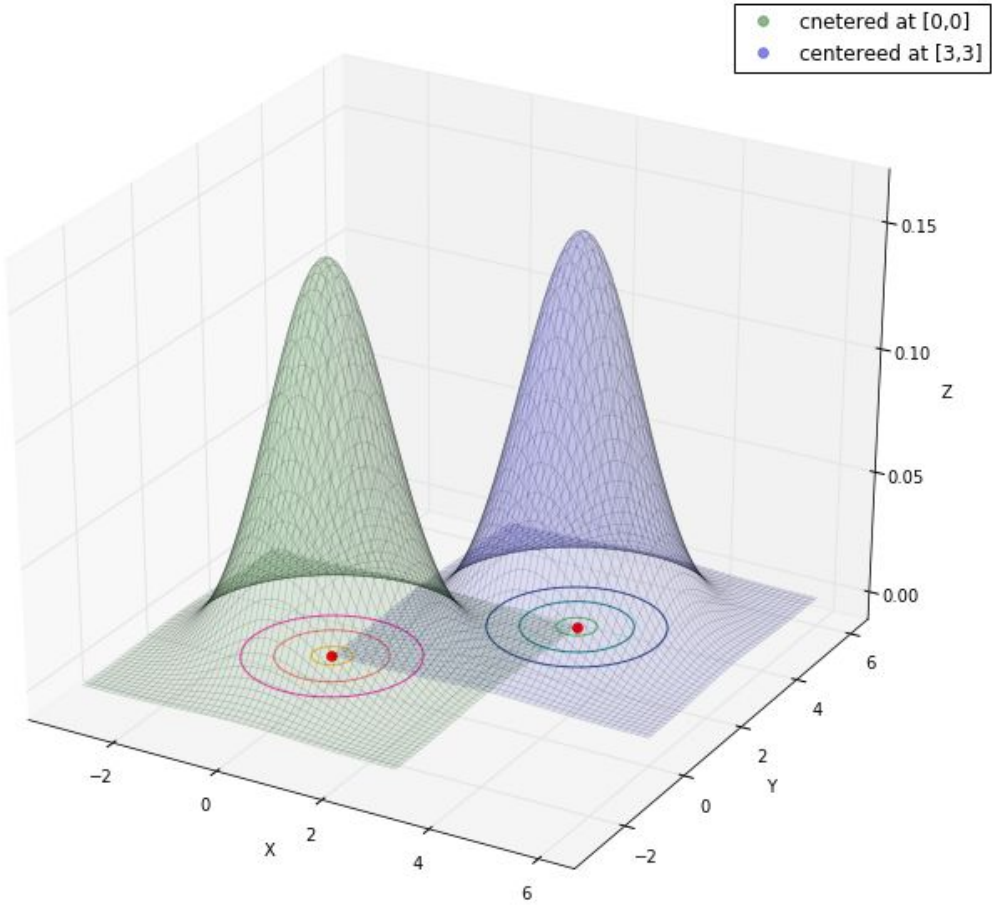
Higher Dimensional Gaussian Distribution

- We still need only "two" parameters to specify higher dimensional Gaussian distribution: the mean and the covariance. However, for a p dimensional case (with p features):
  - the mean is a p-dimensional vector.
  - the covariance is a $p \times p$ matrix.
- The distribution becomes:

$$N(\mu, \Sigma)(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} exp\Big[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\Big]$$

Higher Dimensional Gaussian Distribution

- **mean**: The mean still decides where the "bell" centered at.

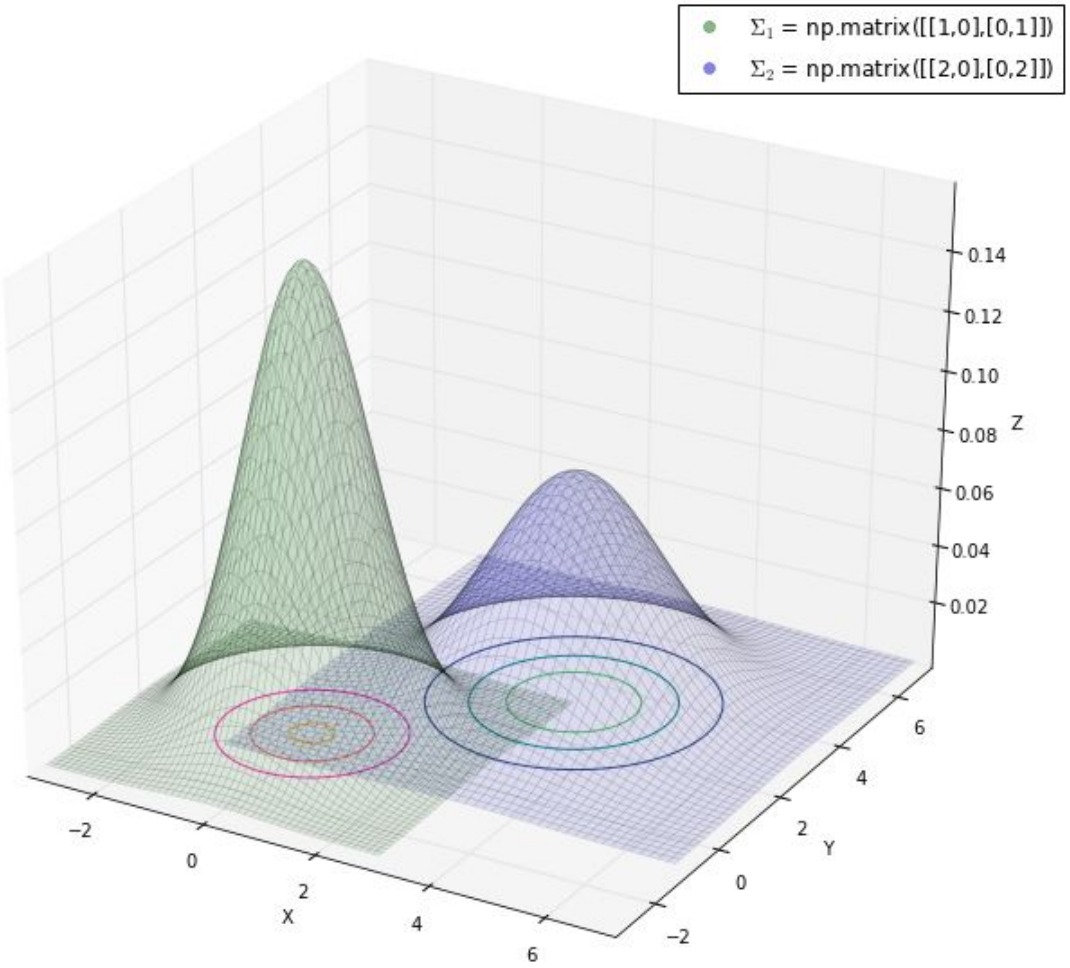


Higher Dimensional Gaussian Distribution

- **Covariance Matrix**: The covariance matrix is a $p \times p$ matrix. The covariance matrix, one of whose special cases is the square of standard deviation in one dimensional space, decide the shape of the "bell". However, the shape means more than just the width in a higher dimensional space.
- **Width**: Let's compare two Gaussian distribution with different covariance matrices in a two dimensional space.

$$\Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ and } \Sigma_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

Higher Dimensional Gaussian Distribution

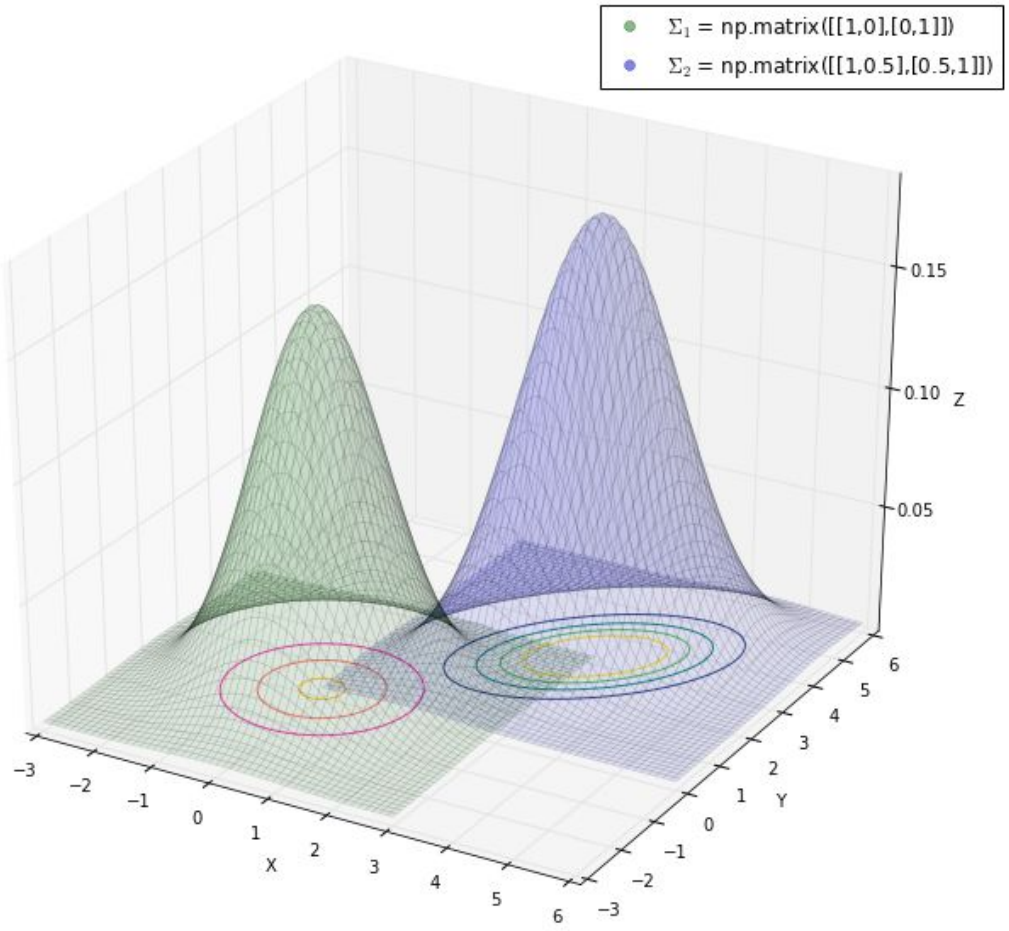


Higher Dimensional Gaussian Distribution

- **Correlation**: Let's compare two Gaussian distribution with different covariance matrices in a two dimensional space.

$$\Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ and } \Sigma_2 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

Higher Dimensional Gaussian Distribution



Models in Higher Dimension

- Recall what we have learned so far:
  - The purpose is to classify based on numerical features, so we want to compute the probability of class k under the condition that X = x:

$$P(Y = k \mid X = x)$$

  - By Bayes theorem we can switch the condition as below:

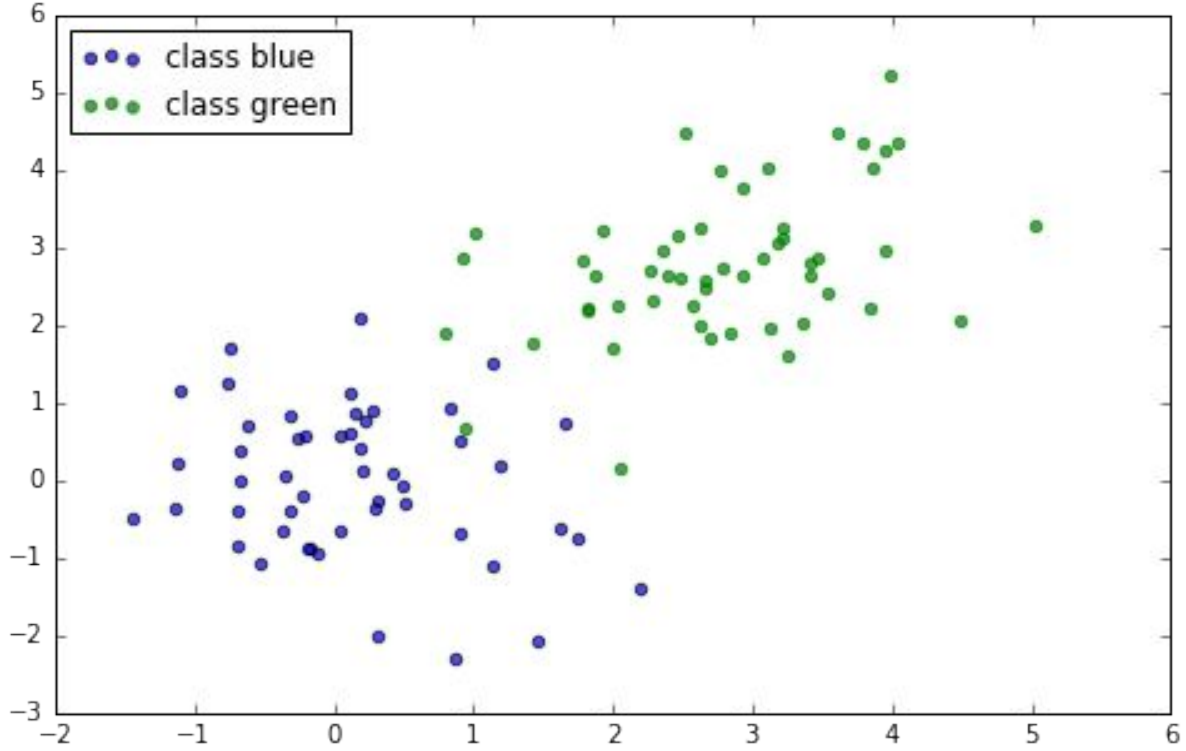
$$P(Y = k \mid X = x) = \frac{p(X = x \mid Y = k)P(Y = k)}{\sum_l p(X = x \mid Y = l)P(Y = l)}$$

Models in Higher Dimension

- So we only need to decide:
  - prior probability.
  - distribution of the feature in each class.
- While estimating prior is easy; estimating the distribution in each class is hard.

Models in Higher Dimension

- To illustrate the difference, let's consider the two-dimensional sample below:



Models in Higher Dimension: QDA

- We start with QDA. For this particular example, we need to compute:

$$p(X = x \mid Y = b) = \frac{1}{(2\pi)|\Sigma_b|^{1/2}} exp\Big[-\frac{1}{2}(x - \mu_b)^T \Sigma_b^{-1}(x - \mu_b)\Big]$$

and

$$p(X = x \mid Y = g) = \frac{1}{(2\pi)|\Sigma_g|^{1/2}} exp\Big[-\frac{1}{2}(x - \mu_g)^T \Sigma_g^{-1}(x - \mu_g)\Big]$$

Models in Higher Dimension: QDA

- With the training data:

$$\text{class blue} = \begin{bmatrix} (x_1^b, y_1^b) \\ (x_2^b, y_2^b) \\ \vdots \\ (x_{50}^b, y_{50}^b) \end{bmatrix} \quad \text{and} \quad \text{class green} = \begin{bmatrix} (x_1^g, y_1^g) \\ (x_2^g, y_2^g) \\ \vdots \\ (x_{53}^g, y_{53}^g) \end{bmatrix}$$

- How do we compute $\hat{\mu}_b$ and $\hat{\mu}_g$?

Models in Higher Dimension: QDA

- The covariance matrices are given by:

$$\hat{\Sigma}_b = \begin{bmatrix} V_x^b & cov_{x,y}^b \\ cov_{y,x}^b & V_y^b \end{bmatrix}$$

What do the parameters in the matrix mean? How do we compute these parameters?

- What about

$$\hat{\Sigma}_g = \begin{bmatrix} V_x^g & cov_{x,y}^g \\ cov_{y,x}^g & V_y^g \end{bmatrix}$$

Models in Higher Dimension: LDA

- We turn to LDA. We need to compute, for each k:

$$p(X = x \mid Y = k) = \frac{1}{(2\pi)|\Sigma|^{1/2}} exp\left[-\frac{1}{2}(x - \mu_k)^T \Sigma^{-1}(x - \mu_k)\right]$$

- **Estimation**
  - We estimate the mean in the same way.
  - We have

$$\hat{\Sigma}^2 = \sum_{k=b,g} \frac{n_k - 1}{n - K} \cdot \hat{\Sigma}_k^2$$

Models in Higher Dimension: GNB

- For GNB, while the mean can be estimated in the same way, we need to estimate the covariance matrix for each k:

$$p(X = x \mid Y = k) = \frac{1}{(2\pi)|\Sigma|^{1/2}} exp\Big[-\frac{1}{2}(x - \mu_k)^T \Sigma^{-1}(x - \mu_k)\Big]$$

- We still have

$$\hat{\Sigma}_b = \begin{bmatrix} V_x^b & cov_{x,y}^b \\ cov_{y,x}^b & V_y^b \end{bmatrix} \qquad \hat{\Sigma}_g = \begin{bmatrix} V_x^g & cov_{x,y}^g \\ cov_{y,x}^g & V_y^g \end{bmatrix}$$

What's the difference (simplification) from QDA?

Models in Higher Dimension: GNB

- GNB assumes

$$\Sigma_b = \begin{bmatrix} V_x^b & 0 \\ 0 & V_y^b \end{bmatrix} \text{ and } \Sigma_g = \begin{bmatrix} V_x^g & 0 \\ 0 & V_y^g \end{bmatrix}$$

- Recall that GNB stands for Gaussian Naive Bayes classifier. The name indicates a Bayes classifier with Gaussian distribution. The assumption of Gaussian distribution, together with uncorrelated features, imply that features are mutually independent. This is what Naive stands for.

Hands-on Session

- Please go to the **"Discriminant Analysis in Scikit-Learn"** in the lecture code.

Outline

- **Limitation of Linear Regression**
- **Logistic Regression**
- **Discriminant Analysis**
  - Linear Discriminant Analysis
  - Quadratic Discriminant Analysis
- **Naive Bayes**

Naive Bayes

- Recall that Bayes theorem assumes the probability that the output is in class *K* given *X* = *x* can be estimated by:

$$Pr(Y = k|X = x) = \frac{f_k(x)\pi_k}{\sum_l f_l(x)\pi_l}$$

- LDA and QDA use multivariate Gaussian densities (but with different assumptions on the covariance matrix). These do not work well when the number of features is large.
- *Naive Bayes* models assume that the inputs are conditionally independent in each class so can work with dataset of large number of features.

Naive Bayes

- The naive Bayes classifier is based on Bayes theorem with independence assumptions between predictors.
- The assumption of conditional independence requires:

$$f_k(x) = \prod_{j=1}^{p} f_{jk}(x)$$

where $f_{jk}(x)$ is the probability density for the j^{th} feature X_j in class k.

- We will introduce three kinds of Naive Bayesian models:
  - Gaussian Naive Bayes
  - Multinomial Naive Bayes
  - Bernoulli Naive Bayes

Gaussian Naive Bayes

- Gaussian Naive Bayes assumes each feature follows a gaussian distribution ($\boldsymbol{\Sigma}_k$ is diagonal):

$$f_{jk}(x) = \frac{1}{\sqrt{2\pi}\sigma_{jk}} exp\Big[-\frac{(x_j - \mu_{jk})^2}{2\sigma_{jk}^2}\Big]$$

  where:

  - μ_{jk}: the mean of the j^{th} feature X_j in class k;
  - σ_{jk}^2: the variance of the j^{th} feature X_j in class k.

- Since we assume gaussian densities, Gaussian Naive Bayes is best used for continuous features.

- `GaussianNB` implements the Gaussian Naive Bayes algorithm for classification.

Hands-on Session

- Please go to the **"Gaussian Naive Bayes in Scikit-Learn"** in the lecture code.

Multinomial Naive Bayes

- If all the features are categorical, then we can parameterize the distribution by vectors $\theta_k = (\theta_{k1}, \theta_{k2}, \ldots, \theta_{kn})$ for each class k, where:
  - n: the number of features.
  - θ_{ki}: probability $P(x_i|k)$ of feature i appearing in a sample belonging to class k.
- `MultinomialNB` implements the naive Bayes algorithm for multinomially distributed data, and is widely used in text classification.

Multinomial Naive Bayes Example

- In our training data, we choose three words to build the model: "sale", "money", "work", denoted by x_1, x_2, x_3.
  - In the spams, "sale" appears 48 times,"money" appears 50 times, "work" 2 times. 100 in total.

    Thus we have:

    - θ_1 = {0.48, 0.50, 0.02}
  - In the non-spams, the frequency of x_1, x_2, x_3 are 5, 10, 85.

    Thus we have:

    - θ_0 = {0.50, 0.10, 0.85}

Hands-on Session

- Please go to the **"Multinomial Naive Bayes in Scikit-Learn"** in the lecture code.

Bernoulli Naive Bayes

- Bernoulli Naive Bayes is used for data that:
  - is distributed according to a multivariate Bernoulli distribution;
  - each feature is assumed to be a binary-valued variable.
- `BernoulliNB` implements the naive Bayes training and classification algorithms.

Bernoulli Naive Bayes

- Consider the spam filter problem. In Bernoulli naive Bayes we do not care about the frequency of a feature. We are just interested in the whether it appears or not.
- Given a feature x_k which denotes a word, does it appears in an email or not? What is the probability of its appearance?

Bernoulli Naive Bayes Example

- Suppose we have 80 non-spams, and the word "sale" (denoted by x_k) appears in 10 of them; we also have 20 spams, and x_k appears in 16 of them. We use $y = 1$ to denote that a email is a spam. Then:

$$p(x_k = 1|y = 1) = \frac{16}{20} = \frac{4}{5}, \quad p(x_k = 0|y = 1) = \frac{1}{5}$$
$$p(x_k = 1|y = 0) = \frac{10}{80} = \frac{1}{8}, \quad p(x_k = 0|y = 0) = \frac{7}{8}$$

- Given a new email which contains the word “sale”, we have class = 1. If we use only this feature to predict:

$$p(y = 1|x_k = 1) = \frac{p(y = 1)p(x_k = 1|y = 1)}{p(x_k = 1)} = \frac{\frac{20}{100} \times \frac{4}{5}}{p(x_k = 1)} = \frac{0.16}{p(x_k = 1)}$$
$$p(y = 0|x_k = 1) = \frac{p(y = 0)p(x_k = 1|y = 0)}{p(x_k = 1)} = \frac{\frac{80}{100} \times \frac{1}{8}}{p(x_k = 1)} = \frac{0.1}{p(x_k = 1)}$$

then we will predict this email to be spam.

Hands-on Session

- Please go to the **"Bernoulli Naive Bayes in Scikit-Learn"** in the lecture code.