

# CH-2-Linear Models for Classification

---

# CONTENTS

- Linear Basis Function Models
- Bayesian Linear Regression
- Discriminant Functions
- Probabilistic Generative Models
- Probabilistic Discriminative Models



# LINEAR MODELS FOR REGRESSION (LINEAR BASIS FUNCTION MODELS )

- Regression-
  - Supervised Learning. Other is Classification.
  - The objective of the regression -to determine the value of one or more of a target variable  $t$ , given the value of a  $D$ -dimensional vector,  $x$  of input variables.
  - Find the function that relates the input and the output.
- Done using Linear Models.
- The polynomial curve fitting is a specific example of a broad class of functions called linear regression models.



# THE LINEAR BASIS FUNCTION

- Given a set of input dataset of  $N$  samples  $\{x_n\}$ , where  $n = 1, \dots, N$ , as well as the corresponding target values  $\{t_n\}$ , the goal is to deduce the value of  $t$  for new value of  $x$ .
- The set of input data set together with the corresponding target values  $t$  is known as the training data set.
- Construct a function  $y(x)$  that maps  $x$  to  $t$  such that:
  - $y(x) = t$  for a new input value of  $x$ .
- Examine this model by finding the probability that the results are correct.
- Examine the probability of  $t$  given  $x$

$$p(t|x)$$



# LINEAR BASIS FUNCTION MODELS

- **Constructing the Linear Basis Function-**

- The basic linear model for regression is a model that involves a linear combination of the input variables:

$$y(x, w) = w_0 + w_1x_1 + w_2x_2 + \dots + w_Dx_D$$

where  $x = (x_1, x_2, \dots, x_D)^T$

- This is what is generally known as linear regression.
- The key attribute of this function-
  - It is a linear function of the parameters  $w_0, w_1, \dots, w_D$  and the input variable  $x_i$ .
  - Being a linear function of the input variable  $x$ , limits the usefulness of the function.
  - Most of the observations that may be encountered does not necessarily follow a linear relationship.
  - To solve this problem consider modifying to model to be a combination of fixed non-linear functions of the input variable.



# THE LINEAR BASIS FUNCTION

- If we assume that the non-linear function of the input variable is  $\varphi(\mathbf{x})$ , then we can re-write the original function as :

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1\varphi(\mathbf{x}_1) + w_2\varphi(\mathbf{x}_2) + \dots + w_D\varphi(\mathbf{x}_D)$$

- Summing it up:

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(x) \leftarrow \text{Linear models}$$

where  $\varphi_j(\mathbf{x})$  are known as basis functions.

$M$  = total number of parameters.

$w_0$  = bias parameter which allows for a fixed offset in the data.



# THE LINEAR BASIS FUNCTION

- convenient to define an additional dummy ‘basis function’  $\phi_0(\mathbf{x}) = 1$  so that

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x})$$

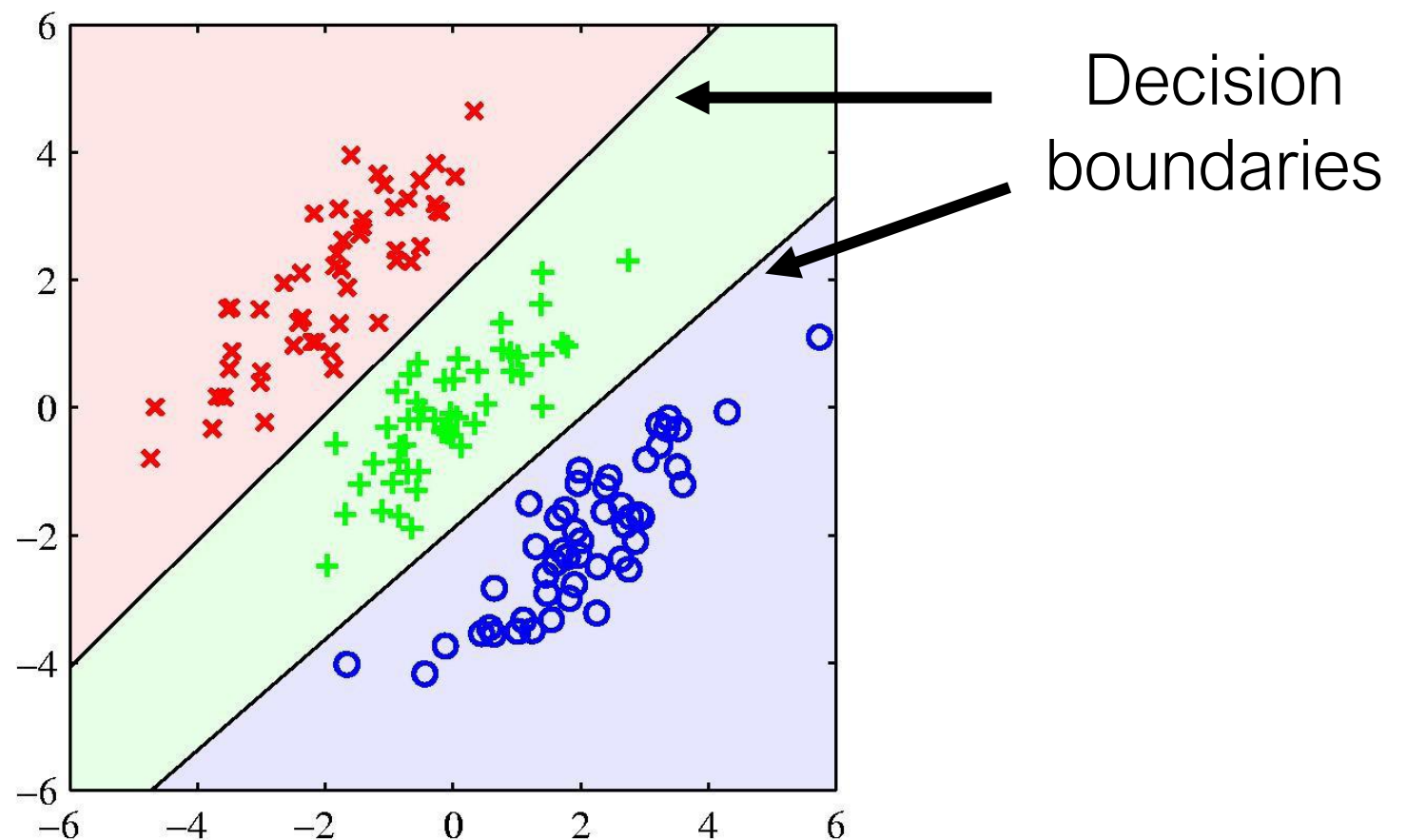
$$\mathbf{w} = (w_0, \dots, w_{M-1})^T \text{ and } \boldsymbol{\phi} = (\phi_0, \dots, \phi_{M-1})^T$$

- In many practical applications of pattern recognition-some form of fixed pre-processing, or feature extraction, to the original data variables
- If the original variables comprise the vector  $\mathbf{x}$ , then the features can be expressed in terms of the basis functions  $\{\phi_j(\mathbf{x})\}$ .

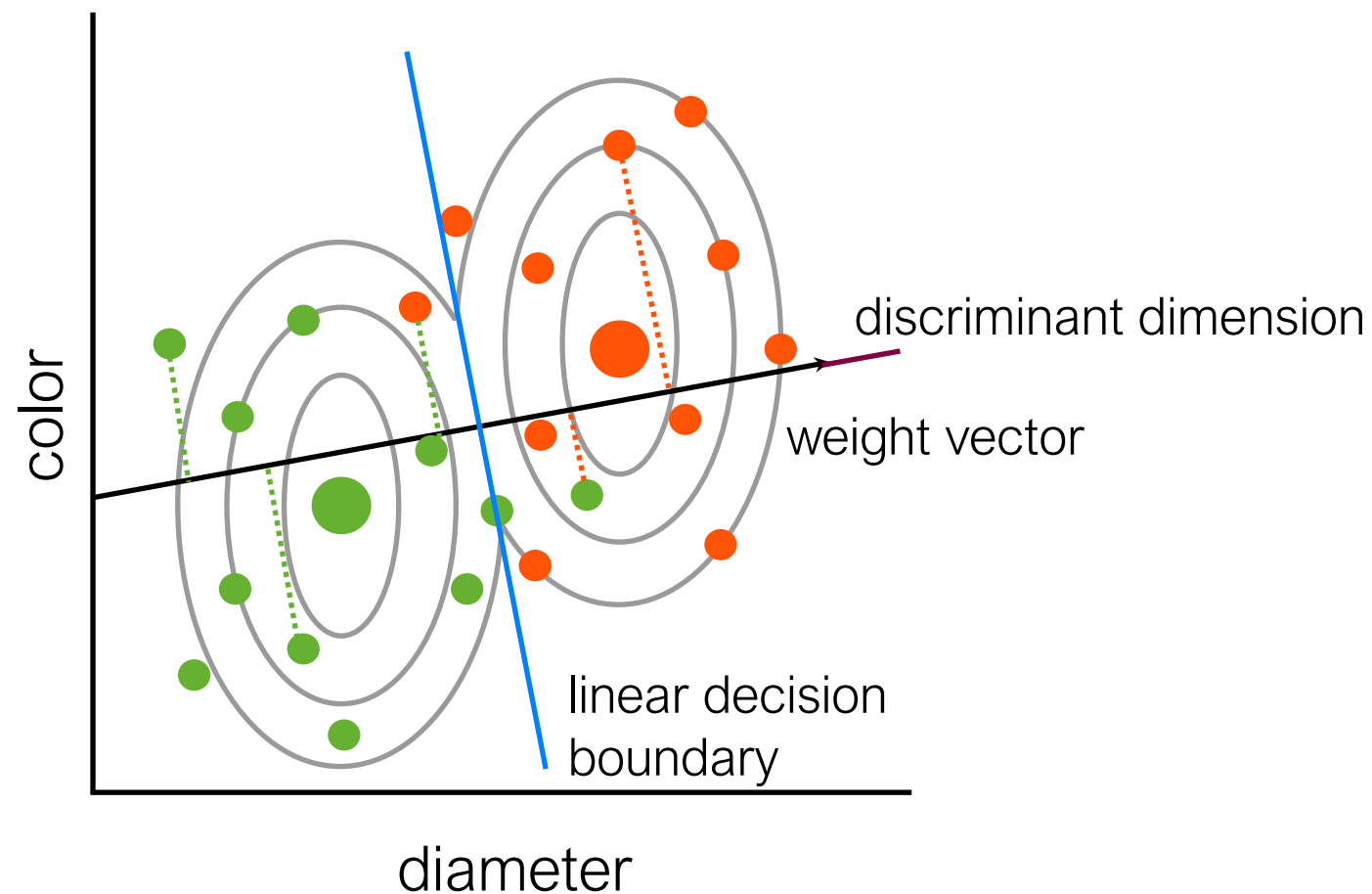


# GOAL

- Our goal is to “classify” input vectors  $\mathbf{x}$  into one of  $k$  classes. Similar to regression, but the output variable is discrete.
- input space is divided into decision regions whose boundaries are called decision boundaries or decision surfaces
- linear models for classification: decision boundaries are linear functions of input vector  $\mathbf{x}$







Classifier seek an 'optimal' separation of classes (e.g., apples and oranges) by finding a set of weights for combining features (e.g., color and diameter).



no computation  
of posterior probabilities  
(probability of certain class given the data)

computation  
of posterior probabilities

Classifier

Discriminant  
function

Probabilistic  
Generative  
Models

Probabilistic  
Discriminative  
Models

- directly map each  $x$  onto a class label

Tools

- Least Square Classification
- Fisher's Linear Discriminant

- model class priors ( $p(C_k)$ ) & class-conditional densities ( $p(x/C_k)$ )
- use to compute posterior probabilities ( $p(C_k/x)$ )

Tools

- Bayes

- model posterior probabilities ( $p(C_k/x)$ ) directly

Tools

- Logistic Regression

# PROS AND CONS OF THE THREE APPROACHES

**Discriminant Functions** are the most simple and intuitive approach to classifying data, but do not allow to

- compensate for class priors (e.g. class 1 is a very rare disease)
- minimize risk (e.g. classifying sick person as healthy more costly than classifying healthy person as sick)
- implement reject option (e.g. person cannot be classified as sick or healthy with a sufficiently high probability)

Probabilistic Generative and Discriminative models can do all that



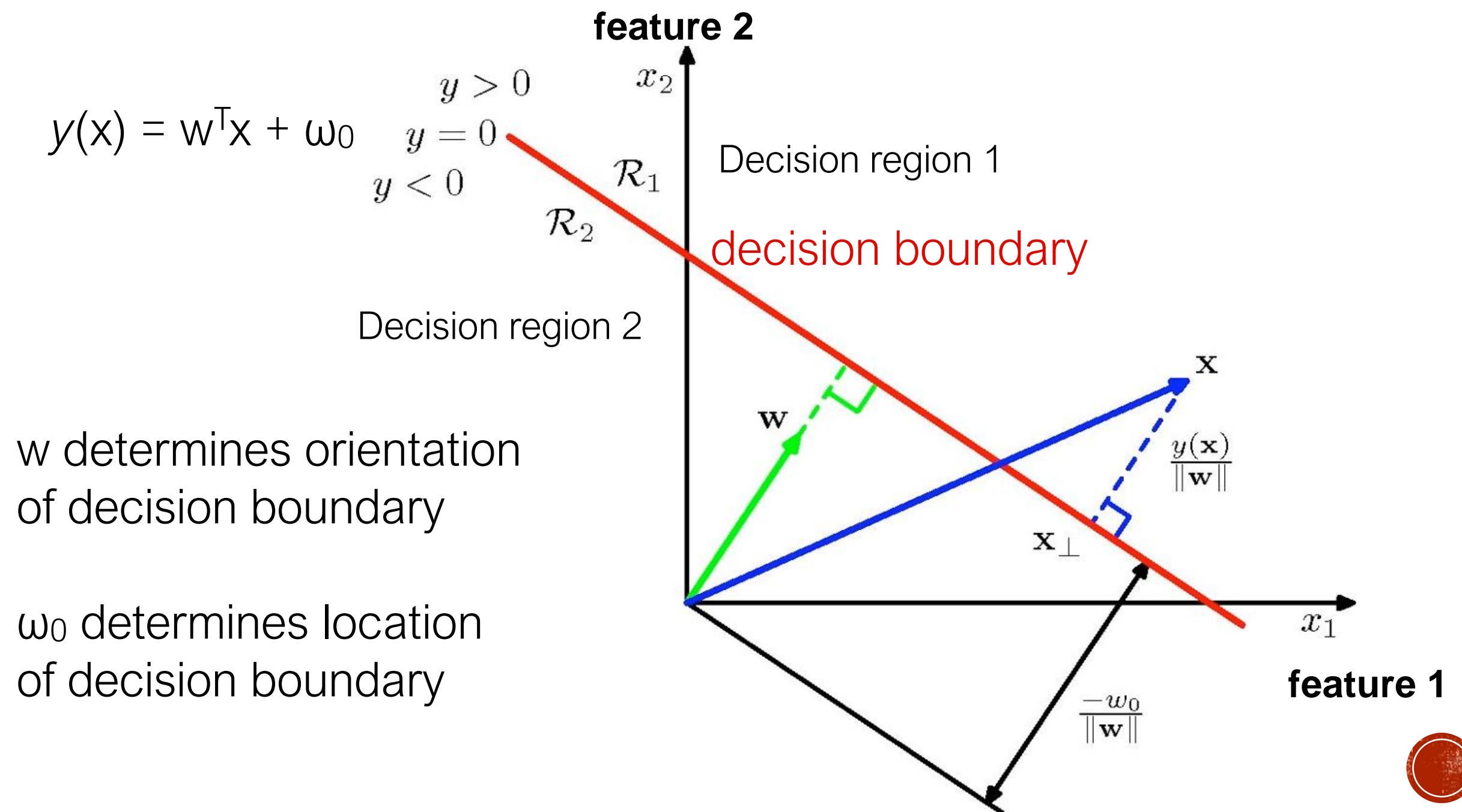
# PROS AND CONS OF THE THREE APPROACHES

- **Generative models**
  - provide a probabilistic model of *all* variables that allows to synthesize new data – but -
  - generating all this information is computationally expensive and complex and is not needed for a simple classification decision
- **Discriminative models**
  - provide a probabilistic model for the target variable (classes) conditional on the observed variables
  - this is usually sufficient for making a well-informed classification decision without the disadvantages of the simple Discriminant Functions



# DISCRIMINANT FUNCTIONS

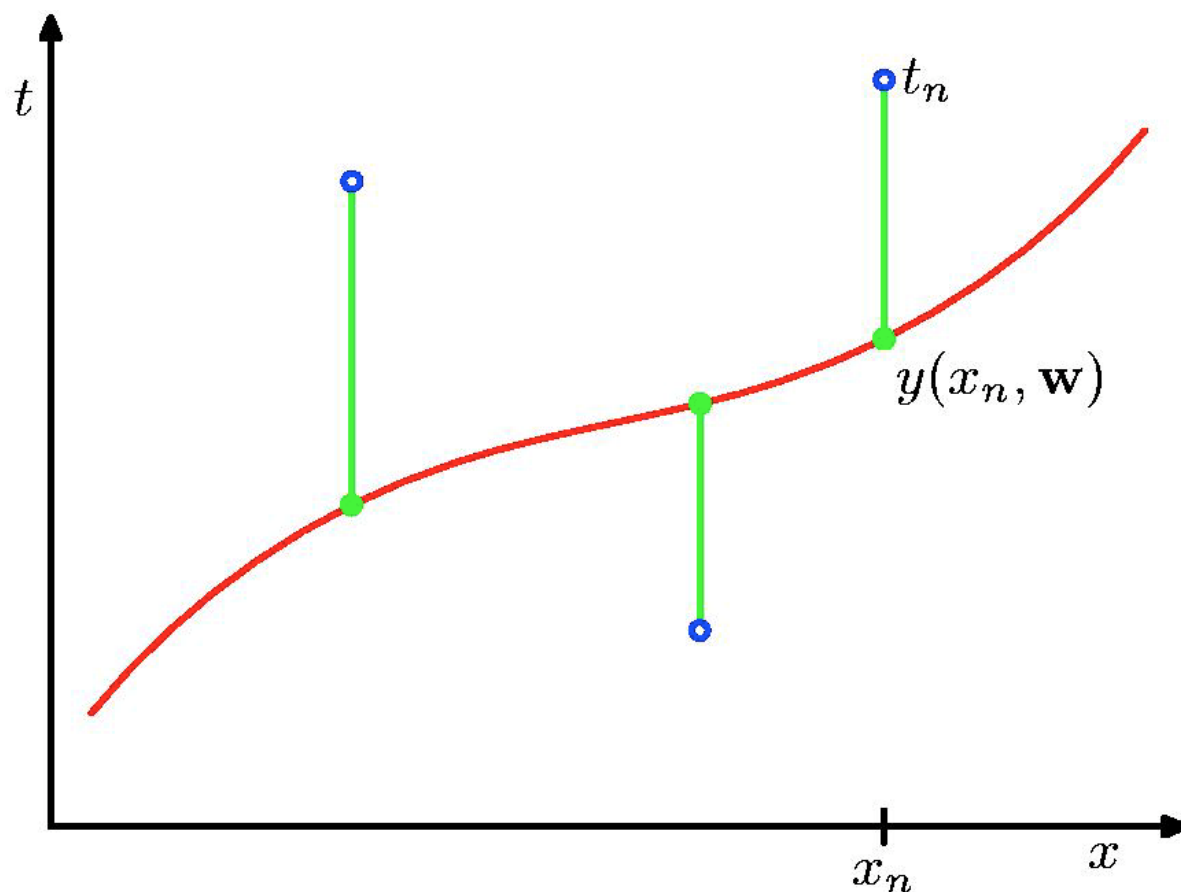
- Functions that are optimized to assign input  $x$  to one of  $k$  classes



# DISCRIMINANT FUNCTIONS - HOW TO DETERMINE PARAMETERS?

## 1. Least Squares for Classification

- General Principle: Minimize the squared distance (residual) between the observed data point and its prediction by a model function

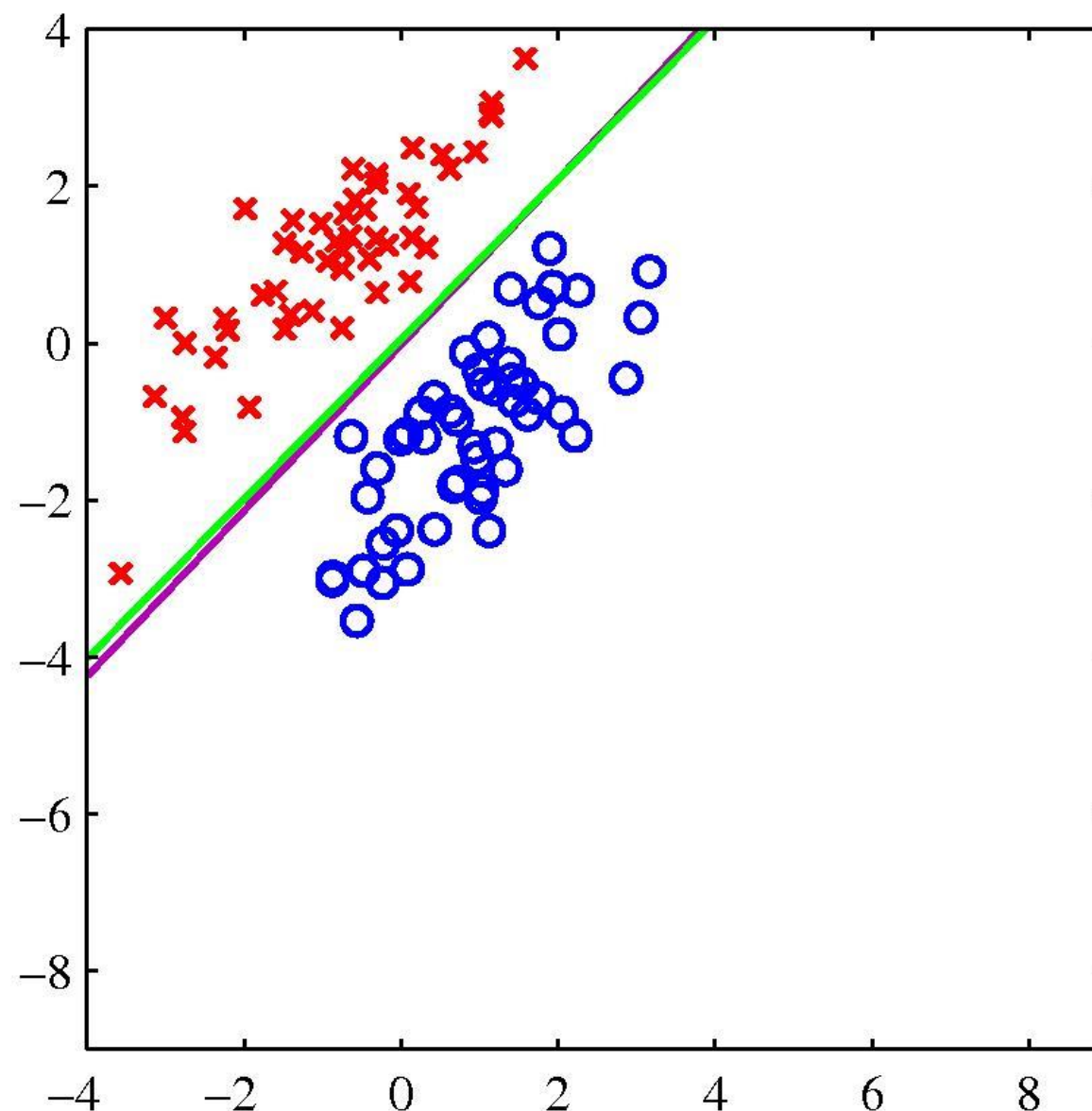


$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2$$



# DISCRIMINANT FUNCTIONS - HOW TO DETERMINE PARAMETERS?

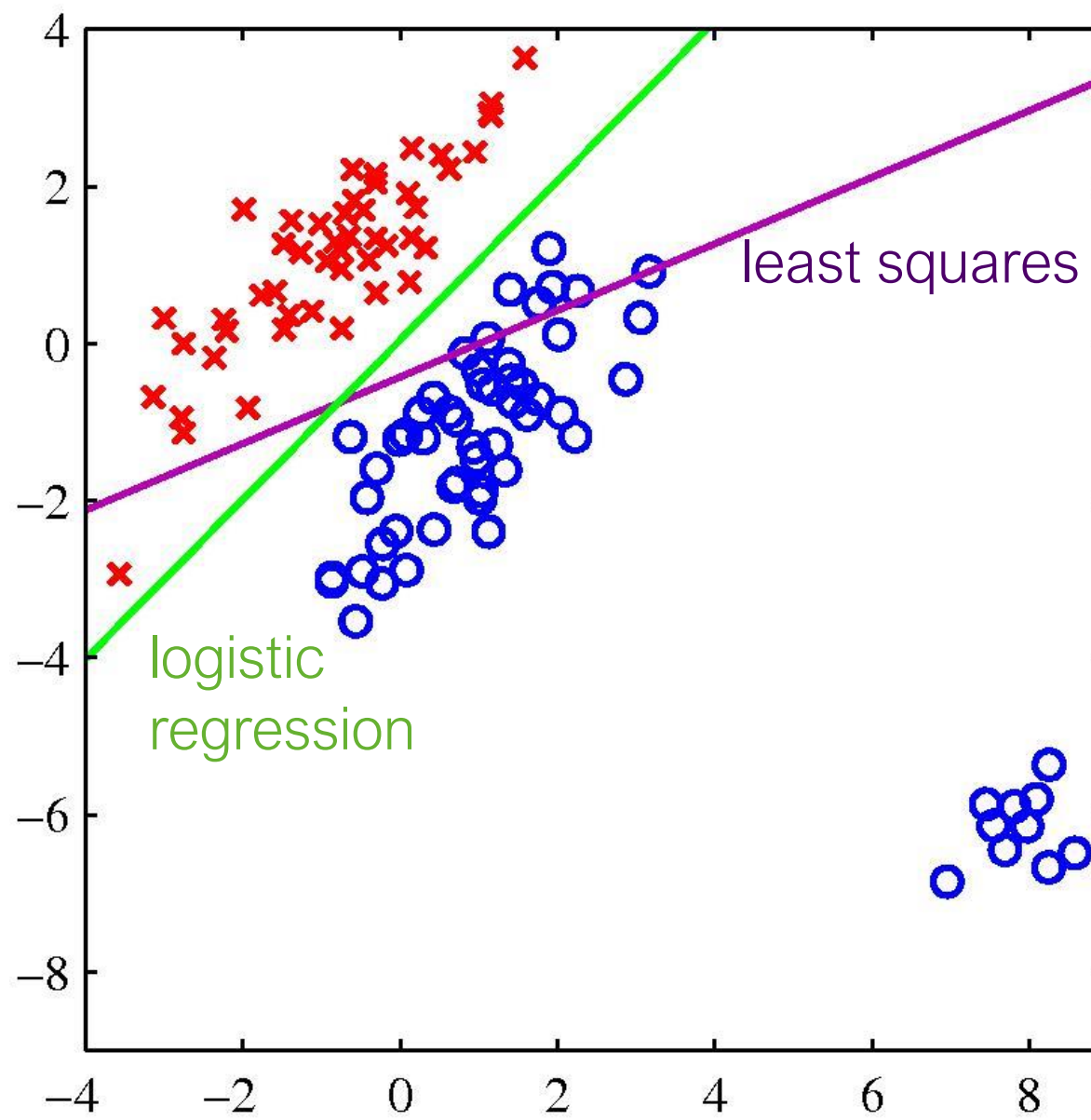
- In the context of classification: find the parameters which minimize the squared distance (residual) between the data points and the decision boundary





# DISCRIMINANT FUNCTIONS - HOW TO DETERMINE PARAMETERS?

- Problem: sensitive to outliers; also distance between the outliers and the discriminant function is minimized --> can shift function in a way that leads to misclassifications





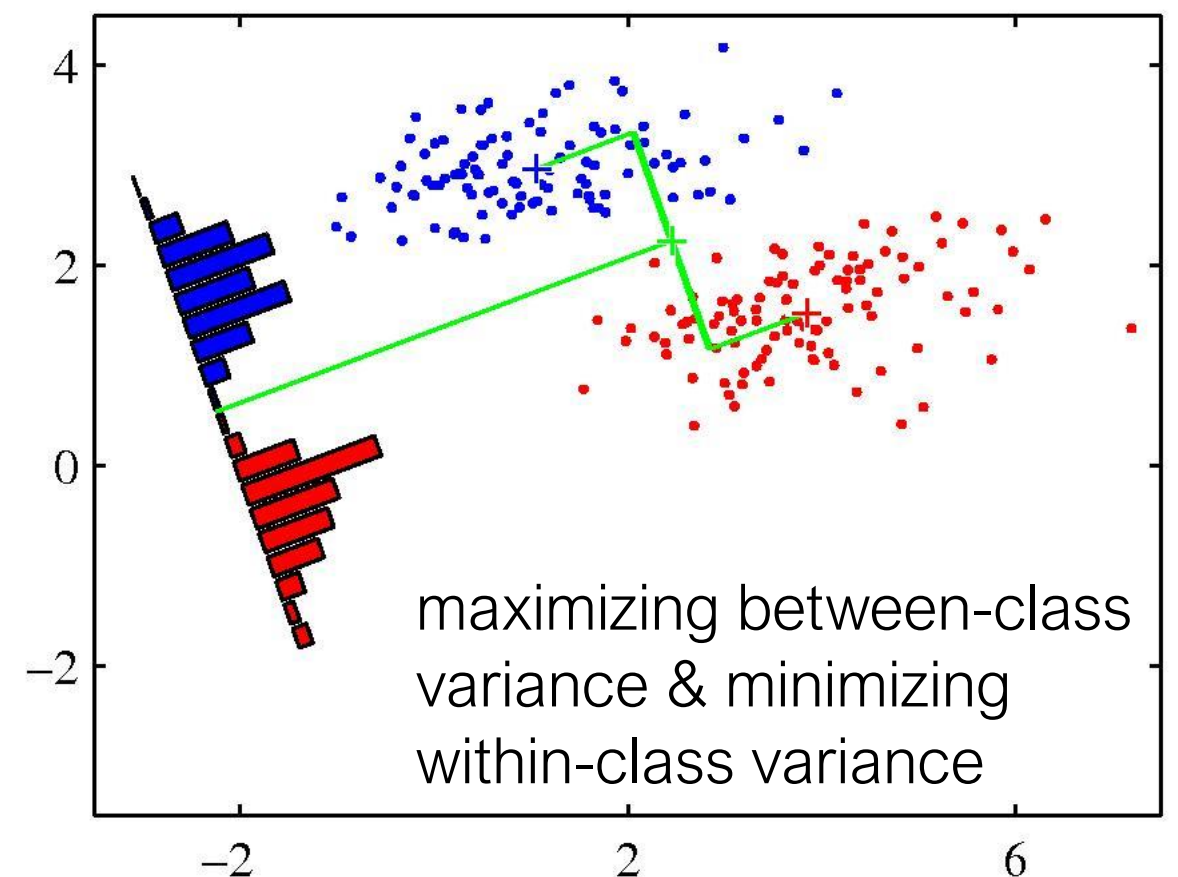
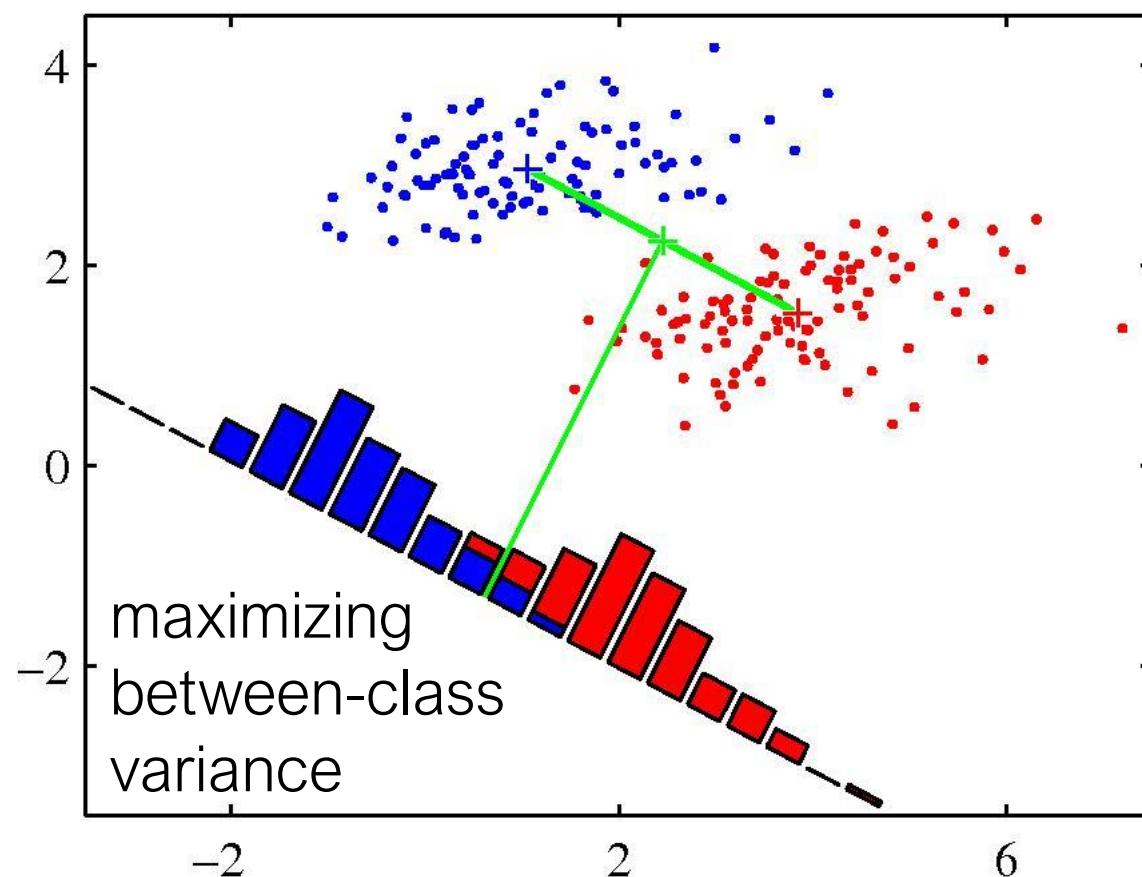
# DISCRIMINANT FUNCTIONS - HOW TO DETERMINE PARAMETERS?

## 2. Fisher's Linear Discriminant

- General Principle: Maximize distance between means of different classes while minimizing the variance within each class

The Fisher criterion

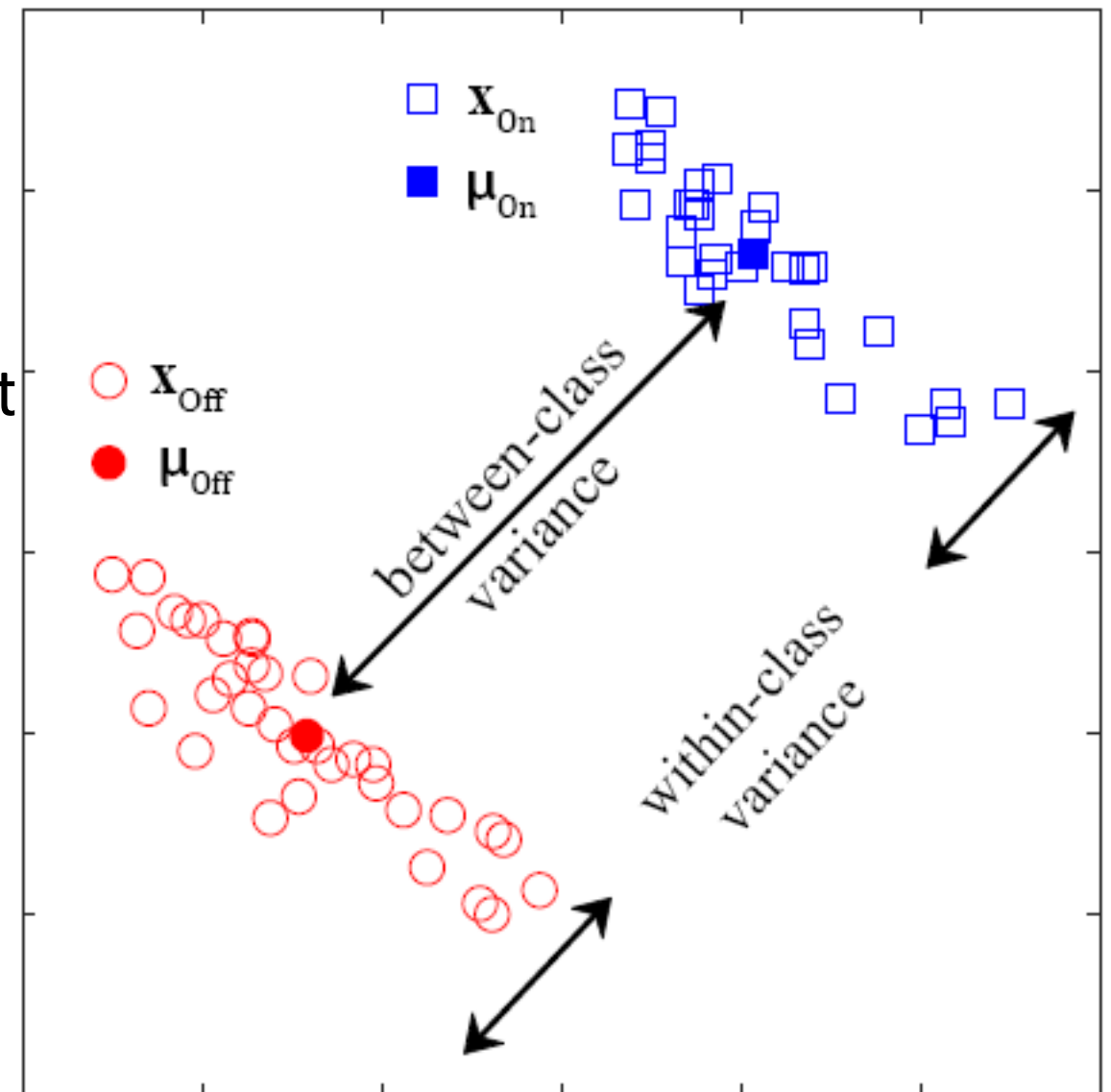
$$J(w) = \frac{|m_1 - m_2|^2}{s_1^2 + s_2^2}$$



# DISCRIMINANT FUNCTIONS - HOW TO DETERMINE PARAMETERS?

## 2. Fisher's Linear Discriminant

- In all problem of data classification, the data would be classified better when the data within each class is close to each other while the data from different class is far to each other.
- Between-class variance (Bvar) and within-class variance (Wvar) are the statistical quantification for these characteristics of data.
- Bvar indicates the separation in the features of two classes while Wvar indicates the scattering of features within each class



# MULTIVARIATE ANALYSIS

- Multiple models
  - Linear regression
  - Logistic regression
  - Cox model
  - Poisson regression
  - Loglinear model
  - Discriminant analysis
  - .....
- Choice of the tool according to the objectives, the study, and the variables



# SIMPLE LINEAR REGRESSION

**Table 1** Age and systolic blood pressure (SBP) among 33 adult women

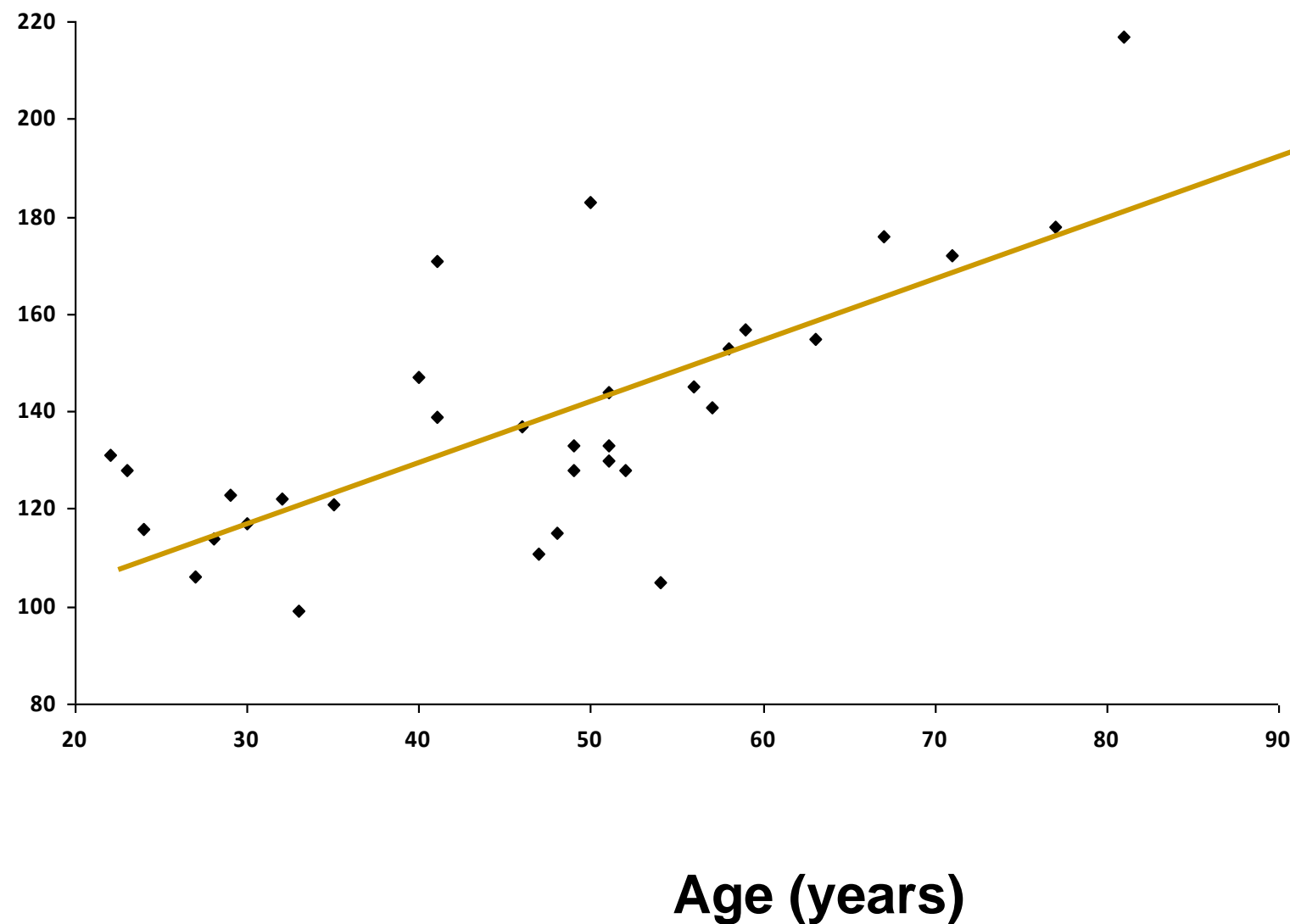
Age	SBP	Age	SBP	Age	SBP
22	131	41	139	52	128
23	128	41	171	54	105
24	116	46	137	56	145
27	106	47	111	57	141
28	114	48	115	58	153
29	123	49	133	59	157
30	117	49	128	63	155
32	122	50	183	67	176
33	99	51	130	71	172
35	121	51	133	77	178
40	147	51	144	81	217



# Simple linear regression

**SBP (mm Hg)**

$$\text{SBP} = 81.54 + 1.222 \cdot \text{Age}$$

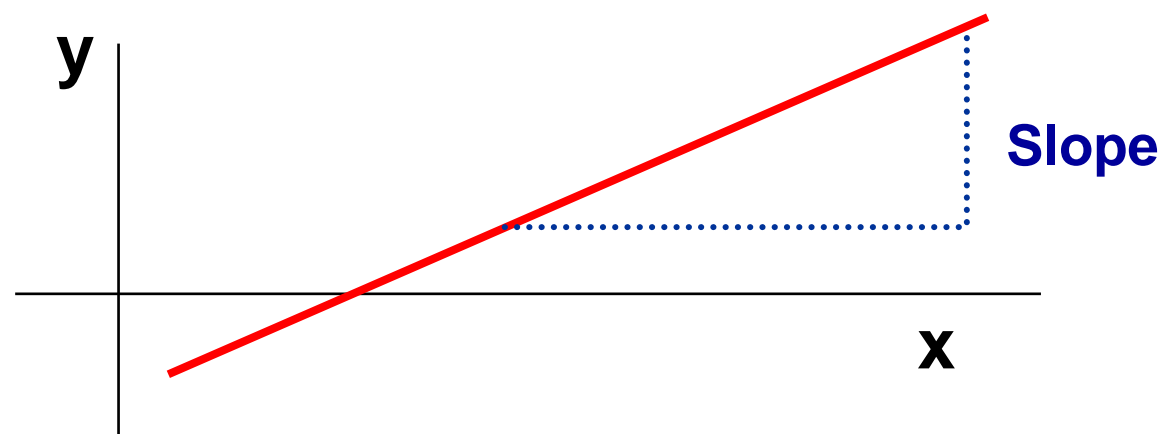


adapted from Colton T. Statistics in Medicine. Boston: Little Brown, 1974



# SIMPLE LINEAR REGRESSION

- Relation between 2 continuous variables (SBP and age)



$$y = \alpha + \beta_1 x_1$$

- Regression coefficient  $\beta_1$ 
  - Measures association between y and x
  - Amount by which y changes on average when x changes by one unit
  - Least squares method



# MULTIPLE LINEAR REGRESSION

- Relation between a continuous variable and a set of  $i$  continuous variables

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_i x_i$$

- Partial regression coefficients  $\beta_i$ 
  - Amount by which  $y$  changes on average when  $x_i$  changes by one unit and all the other  $x_i$ s remain constant
  - Measures association between  $x_i$  and  $y$  adjusted for all other  $x_i$
- Example
  - SBP *versus* age, weight, height, etc



# MULTIPLE LINEAR REGRESSION

$$\underline{y} = \underline{\alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_i x_i}$$

Predicted

Response variable

Outcome variable

Dependent

Predictor variables

Explanatory variables

Covariables

Independent variables





# LOGISTIC REGRESSION

**Table 2 Age and signs of coronary heart disease (CD)**

Age	CD	Age	CD	Age	CD
22	0	40	0	54	0
23	0	41	1	55	1
24	0	46	0	58	1
27	0	47	0	60	1
28	0	48	0	60	0
30	0	49	1	62	1
30	0	49	0	65	1
32	0	50	1	67	1
33	0	51	0	71	1
35	1	51	1	77	1
38	0	52	0	81	1

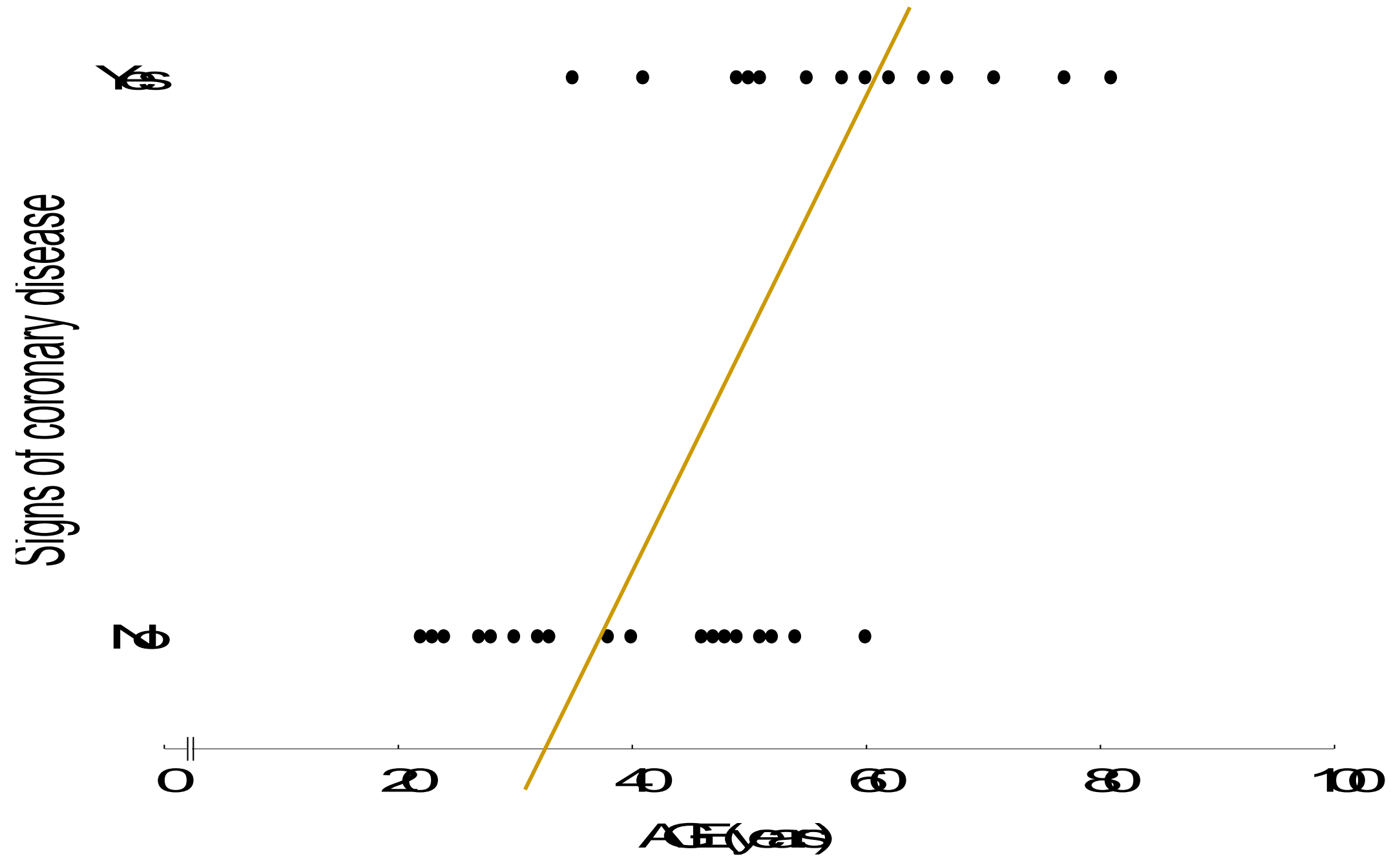


# HOW CAN WE ANALYSE THESE DATA?

- Compare mean age of diseased and non-diseased
  - Non-diseased: 38.6 years
  - Diseased: 58.7 years ( $p < 0.0001$ )
- Linear regression?



# DOT-PLOT: DATA FROM TABLE 2



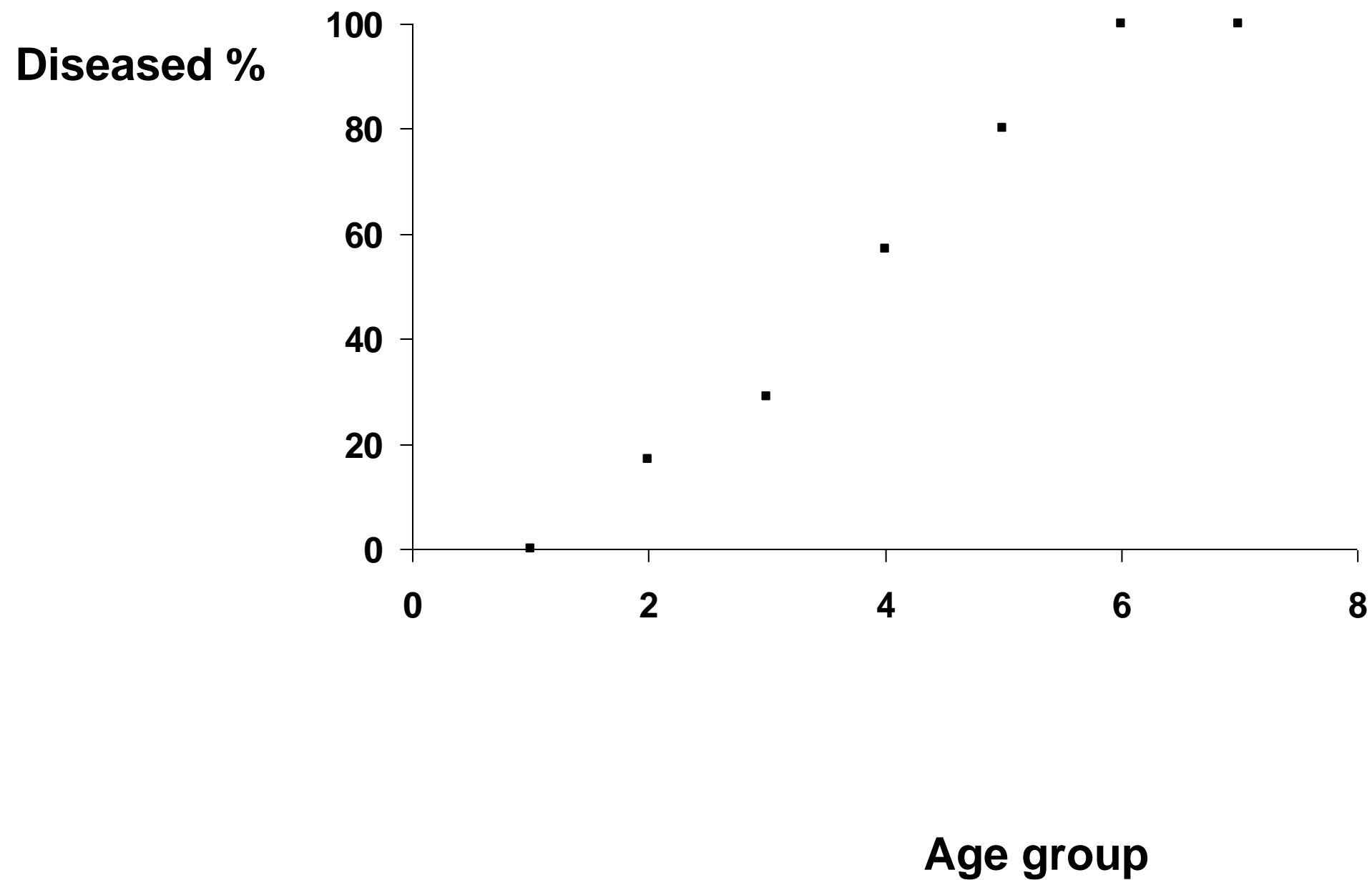
# LOGISTIC REGRESSION (2)

Table 3 Prevalence (%) of signs of CD according to age group

Age group	# in group	Diseased	
		#	%
20 - 29	5	0	0
30 - 39	6	1	17
40 - 49	7	2	29
50 - 59	7	4	57
60 - 69	5	4	80
70 - 79	2	2	100
80 - 89	1	1	100



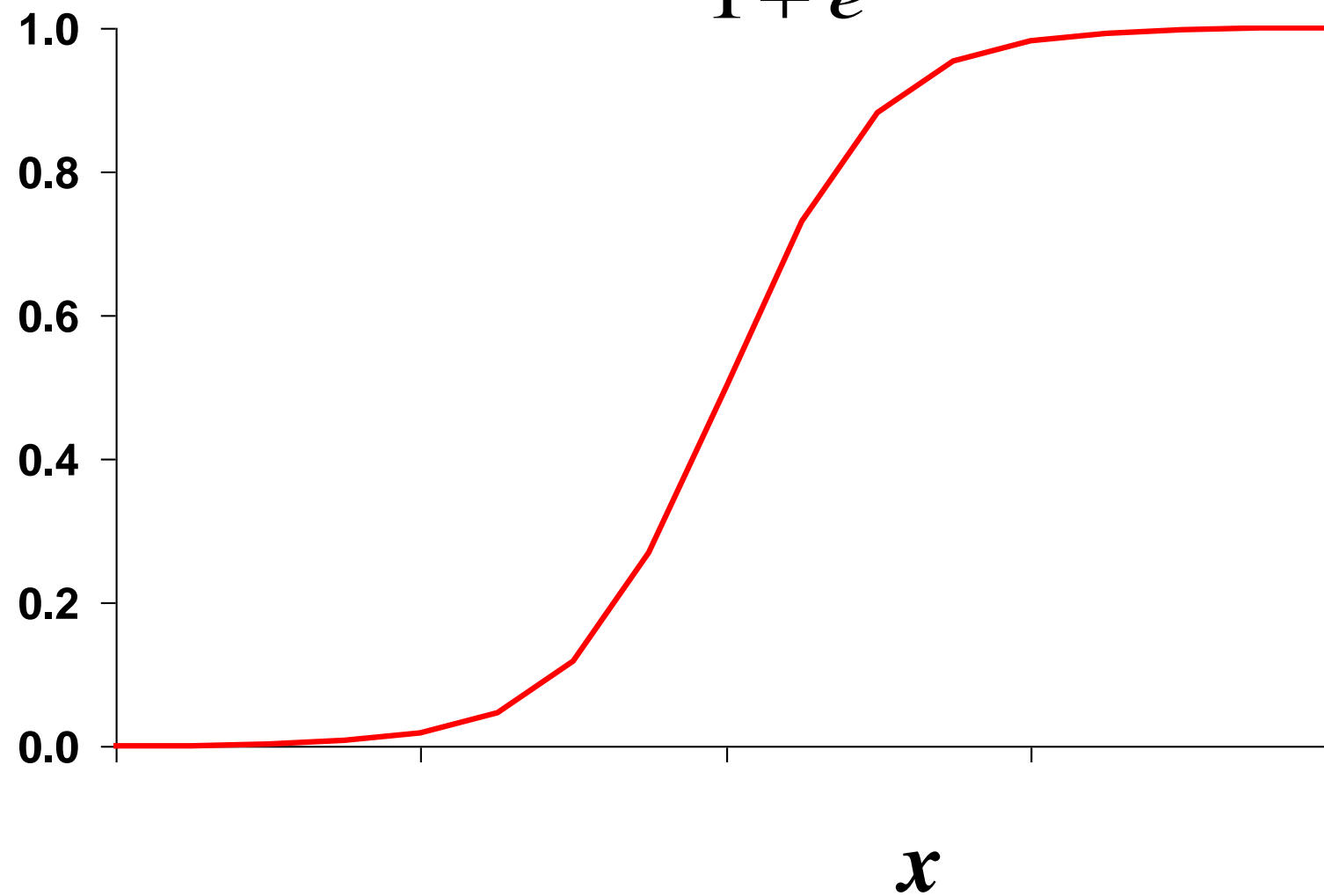
# DOT-PLOT: DATA FROM TABLE 3



# LOGISTIC FUNCTION

$$P(y|x) = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}$$

Probability of  
disease



# TRANSFORMATION

$$P(y|x) = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}$$

$$\frac{P(y|x)}{1 - P(y|x)}$$

$$\ln \left[ \frac{P(y|x)}{1 - P(y|x)} \right] = \alpha + \beta x$$



logit of  $P(y/x)$

✓  $\alpha$  = log odds of disease  
in unexposed

✓  $\beta$  = log odds ratio associated  
with being exposed

✓  $e^{\beta}$  = odds ratio



# DISCRIMINATIVE CLASSIFIER: LOGISTIC REGRESSION

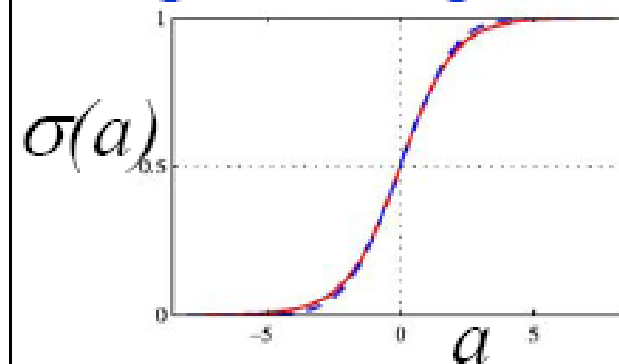
- Feature vector  $\mathbf{x}$
- Two-class classification: class variable  $y$  has values  $C_1$  and  $C_2$
- *A posteriori* probability  $p(C_1|\mathbf{x})$  written as

$$p(C_1|\mathbf{x}) = f(\mathbf{x}) = \sigma(\mathbf{w}^T \mathbf{x}) \text{ where}$$

$$\sigma(a) = \frac{1}{1 + \exp(-a)}$$

- It is known as logistic regression in statistics
  - Although it is a model for classification rather than for regression

## Logistic Sigmoid



### Properties:

#### A. Symmetry

$$\sigma(-a) = 1 - \sigma(a)$$

#### B. Inverse

$$a = \ln(\sigma / (1 - \sigma))$$

known as *logit*.  
Also known as *log odds* since it is the ratio

$$\ln[p(C_1|x)/p(C_2|x)]$$

#### C. Derivative

$$d\sigma/da = \sigma(1 - \sigma)$$





# FITTING EQUATION TO THE DATA

- Linear regression: Least squares
- Logistic regression: Maximum likelihood
- Likelihood function
  - Estimates parameters  $\alpha$  and  $\beta$
  - Practically easier to work with log-likelihood

$$L(B) = \ln[l(B)] = \sum_{i=1}^n \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\}$$



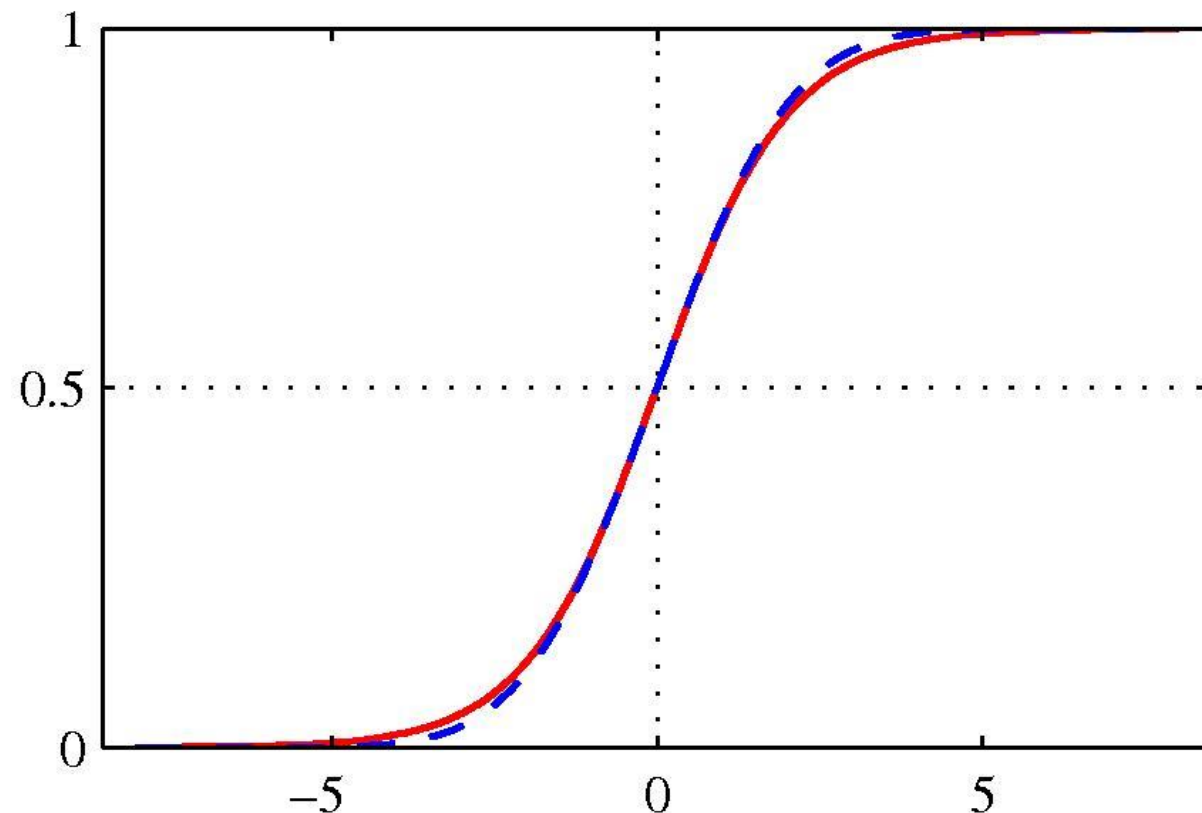
# MAXIMUM LIKELIHOOD

- Iterative computing
  - Choice of an arbitrary value for the coefficients (usually 0)
  - Computing of log-likelihood
  - Variation of coefficients' values
  - Reiteration until maximisation (plateau)
- Results
  - Maximum Likelihood Estimates (MLE) for  $\alpha$  and  $\beta$
  - Estimates of  $P(y)$  for a given value of  $x$



# PROBABILISTIC GENERATIVE MODELS

- model class-conditional densities ( $p(x \mid C_k)$ ) and class priors ( $p(C_k)$ )
- use them to compute posterior class probabilities ( $p(C_k \mid x)$ ) according to Bayes theorem
- posterior probabilities can be described as logistic sigmoid function



inverse of sigmoid function is the logit function which represents the ratio of the posterior probabilities for the two classes

$$\ln[p(C_1 \mid x)/p(C_2 \mid x)] \rightarrow \text{log odds}$$



# PROBABILISTIC DISCRIMINATIVE MODELS - LOGISTIC REGRESSION

- An example of a *probabilistic discriminative model*
- Rather than learning  $P(\mathbf{x}|C_i)$  and  $P(C_i)$ , attempts to directly learn  $P(C_i|\mathbf{x})$
- Advantages: fewer parameters, better if assumptions in class-conditional density formulation are inaccurate
- We have seen how the class posterior for a two-class setting can be written as a logistic sigmoid acting on a linear function of the feature vector  $\boldsymbol{\phi}$ :

$$p(C_1|\boldsymbol{\phi}) = y(\boldsymbol{\phi}) = \sigma(\mathbf{w}^T \boldsymbol{\phi})$$

- This model is called *logistic regression*, even though it is a model for **classification**, not regression!



# PROBABILISTIC DISCRIMINATIVE MODELS - LOGISTIC REGRESSION

- you model the posterior probabilities directly assuming that they have a sigmoid-shaped distribution (without modeling class priors and class-conditional densities)
- the sigmoid-shaped function ( $\sigma$ ) is model function of logistic regressions
- first non-linear transformation of inputs using a vector of basis functions  $\phi(x) \rightarrow$  suitable choices of basis functions can make the modeling of the posterior probabilities easier

$$p(C_1/\phi) = y(\phi) = \sigma(w^T \phi)$$

$$p(C_2/\phi) = 1 - p(C_1/\phi)$$



# PROBABILISTIC DISCRIMINATIVE MODELS - LOGISTIC REGRESSION

- Parameters of the logistic regression model determined by maximum likelihood estimation
- maximum likelihood estimates are computed using iterative reweighted least squares → iterative procedure that minimizes error function using mathematical algorithms (Newton-Raphson iterative optimization scheme)
- that means starting from some initial values the weights are changed until the likelihood is maximized



# LOGISTIC REGRESSION

**Data:** Inputs are continuous vectors of length  $M$ . Outputs are discrete.

$$\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^N \text{ where } \mathbf{x} \in \mathbb{R}^M \text{ and } y \in \{0, 1\}$$

**Model:** Logistic function applied to dot product of parameters with input vector.

$$p_{\boldsymbol{\theta}}(y = 1|\mathbf{x}) = \frac{1}{1 + \exp(-\boldsymbol{\theta}^T \mathbf{x})}$$

**Learning:** finds the parameters that minimize some objective function.  $\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} J(\boldsymbol{\theta})$

**Prediction:** Output is the most probable class.

$$\hat{y} = \underset{y \in \{0,1\}}{\operatorname{argmax}} p_{\boldsymbol{\theta}}(y|\mathbf{x})$$



# Probabilistic Discriminative Models

Two-class case:  $p(C_1 | \mathbf{x}) = \sigma(\mathbf{w}^T \mathbf{x} + w_0)$

Multiclass case: 
$$p(C_k | \mathbf{x}) = \frac{e^{\mathbf{w}_k^T \mathbf{x} + w_{k0}}}{\sum_j e^{\mathbf{w}_j^T \mathbf{x} + w_{j0}}}$$

Discriminative approach: use the functional form of the generalized linear model for the posterior probabilities and determine its parameters directly using maximum likelihood.





# Probabilistic Discriminative Models

## Advantages:

- Fewer parameters to be determined
- Improved predictive performance, especially when the class-conditional density assumptions give a poor approximation of the true distributions.



# Probabilistic Discriminative Models

Two-class case:

$$p(C_1 | \mathbf{x}) = \sigma(\mathbf{w}^T \mathbf{x} + w_0) = y(\mathbf{x})$$

$$p(C_2 | \mathbf{x}) = 1 - p(C_1 | \mathbf{x})$$

In the terminology of statistics, this model is known as logistic regression.

Assuming  $\mathbf{x} \in \mathbb{R}^M$  how many parameters do we need to estimate?

$$M + 1$$

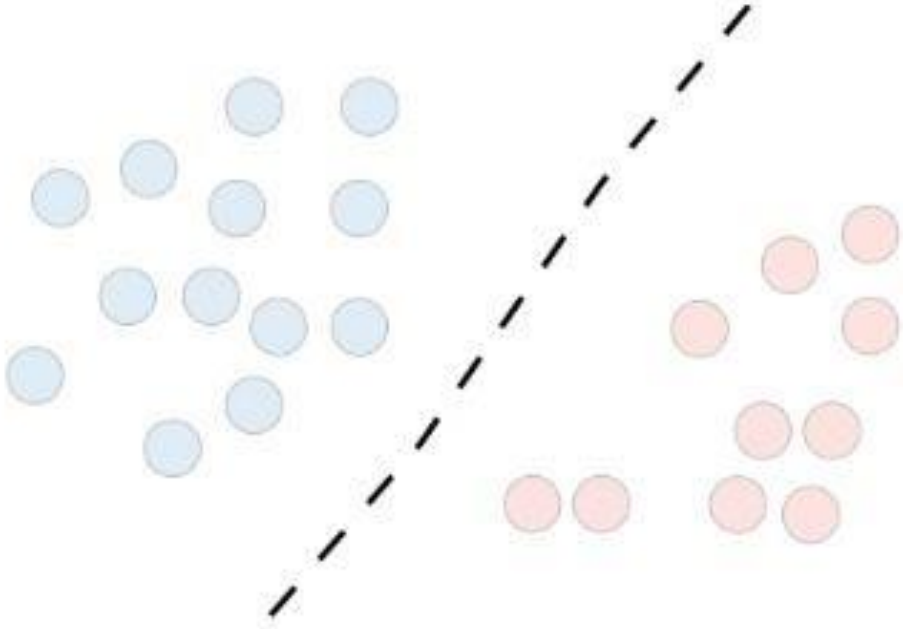
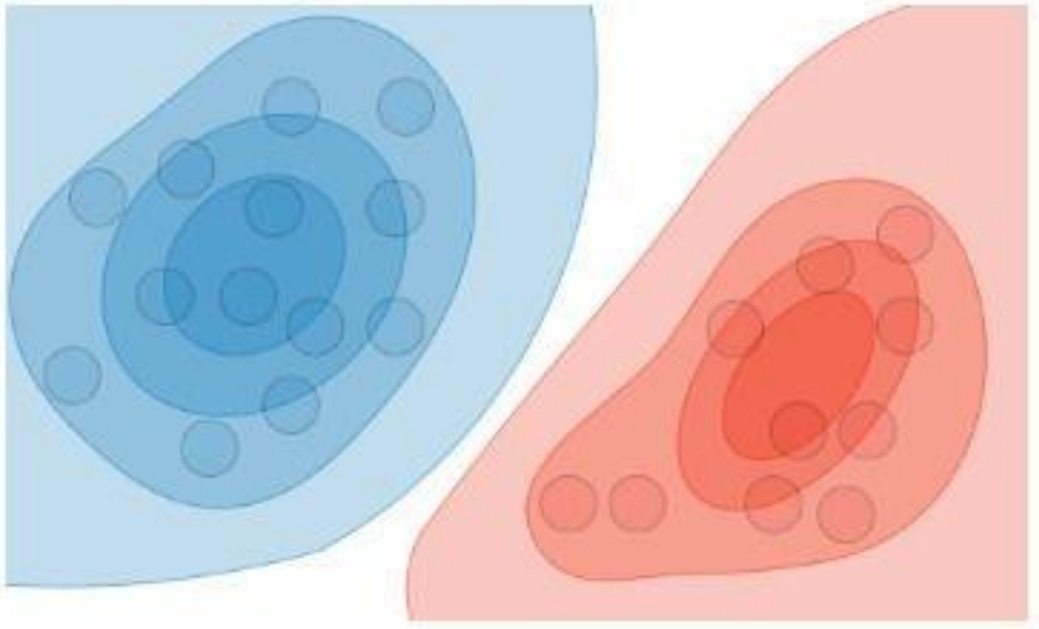


# DISCRIMINATIVE VS GENERATIVE:

	Generative Model	Discriminative Model
Learns	Probabilistic model	Decision boundary
Estimates	$P(x, y)$	$P(y   x)$
Strength	Converges faster	Smaller error
Explainability	Express complex relationships	Low to none
Examples	Naive Bayes Classifier, GAN	Linear Regression, SVM



# DISCRIMINATIVE VS GENERATIVE:

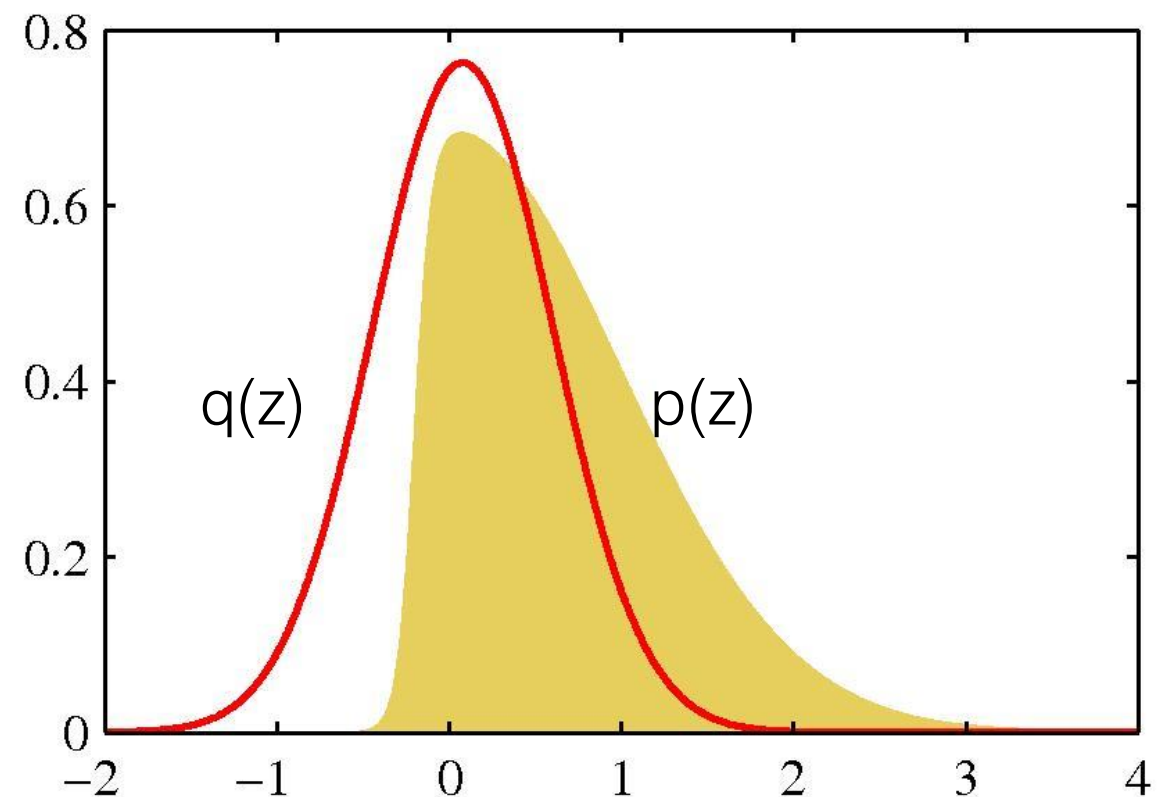
	Discriminative model	Generative model
Goal	Directly estimate $P(y x)$	Estimate $P(x y)$ to then deduce $P(y x)$
What's learned	Decision boundary	Probability distributions of the data
Illustration		
Examples	Regressions, SVMs	GDA, Naive Bayes



# NORMALIZING POSTERIOR PROBABILITIES

- To compare models and to use posterior probabilities in Bayesian Logistic Regression it is useful to have posterior probabilities in Gaussian form
- LAPLACE APPROXIMATION is the tool to find a Gaussian approximation to a probability density defined over a set of continuous variables; here it is used to find a gaussian approximation of your posterior probabilities
- Goal is to find Gaussian approximation  $q(z)$  centered on the mode of  $p(z)$

$$p(z) = 1/Z f(z) \quad \begin{array}{l} Z = \text{unknown} \\ \text{normalization constant} \end{array}$$



# HOW TO FIND THE BEST MODEL? - BAYES INFORMATION CRITERION (BIC)

- the approximation of the normalization constant  $Z$  can be used to obtain an approximation for the model evidence
- Consider data set  $D$  and models  $\{M_i\}$  having parameters  $\{\theta_i\}$
- For each model define likelihood  $p(D|\theta_i, M_i)$
- Introduce prior over parameters  $p(\theta_i|M_i)$
- Need model evidence  $p(D|M_i)$  for various models
- $Z$  is approximation of model evidence  $p(D|M_i)$



# MAKING PREDICTIONS

- having obtained a Gaussian approximation of your posterior distribution (using Laplace approximation) you can make predictions for new data using BAYESIAN LOGISTIC REGRESSION
- you use the normalized posterior distribution to arrive at a predictive distribution for the classes given new data
- you marginalize with respect to the normalized posterior distribution



# TYPES OF LINEAR REGRESSION

- **Simple linear regression:** This involves modeling the relationship between a single input variable (explanatory variable) and a single output variable (response variable). The model is represented by a straight line, and the goal is to find the line that best fits the data.
- **Multiple linear regression:** This involves modeling the relationship between multiple input variables and a single output variable. The model is represented by a straight line, and the goal is to find the line that best fits the data.
- **Polynomial regression:** This involves modeling the relationship between an input variable and an output variable using a polynomial function. The model is represented by a curve, and the goal is to find the curve that best fits the data.





# TYPES OF LINEAR REGRESSION

- **Logistic regression:** This is a type of regression used when the output variable is binary (e.g., 0 or 1, Yes or No). The model is used to predict the probability that a given input belongs to one of the two categories.
- **Ridge regression:** This is a variation of multiple linear regression that adds a penalty term to the objective function to discourage the model from overfitting the data.
- **Lasso regression:** This is another variation of multiple linear regression that adds a penalty term to the objective function to discourage the model from overfitting the data. Unlike ridge regression, lasso regression can zero out some of the coefficients, effectively removing some of the input variables from the model.



# LINEAR REGRESSION-EXAMPLE

SUBJECT	AGE X	GLUCOSE LEVEL Y
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81
7	55	?



# LINEAR REGRESSION-EXAMPLE

The key steps for regression are simple:

1. List all the variables available for making the model.
2. Establish a dependent variable of interest.
3. Examine visual (if possible) relationships between variables of interest.
4. Find a way to predict the dependent variables using the other variables.

Estimated  
(or predicted)  
Y value for  
observation i

Estimate of  
the regression  
intercept

Estimate of the  
regression slope

Value of X for  
observation i

$$\hat{Y}_i = b_0 + b_1 X_i$$



# LINEAR REGRESSION-EXAMPLE

$$b_0 = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

$$b_1 = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$



# LINEAR REGRESSION-EXAMPLE

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X <sup>2</sup>	Y <sup>2</sup>
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561
$\Sigma$	247	486	20485	11409	40022



# LINEAR REGRESSION-EXAMPLE

Find  $b_0$ :

$$b_0 = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

$$b_0 = \frac{(486)(11409) - (247)(20485)}{6(11409) - (247)^2}$$

$$b_0 = \frac{4848979}{7445} = 65.14$$



# LINEAR REGRESSION-EXAMPLE

Find  $b_1$ :

$$b_1 = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

$$b_1 = \frac{6(20485) - (247)(486)}{6(11409) - (247)^2}$$

$$b_1 = \frac{2868}{7445} = 0.385335$$



# LINEAR REGRESSION-EXAMPLE

**Step 3:** *Insert the values into the equation.*

$$y' = b_0 + b_1 * x$$

$$y' = 65.14 + (0.385225 * x)$$

**Step 4:** *Prediction – the value of y for the given value of x = 55*

$$y' = 65.14 + (0.385225 * 55)$$

$$y' = 86.327$$





# LINEAR REGRESSION-EXAMPLE

- Find linear regression equation for the following two sets of data:

x	2	4	6	8
y	3	7	5	10



# LINEAR REGRESSION-EXAMPLE

$$b = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$

$$b = \frac{4 \times 144 - 20 \times 25}{4 \times 120 - 400}$$

$$b = 0.95$$

$$a = \frac{\sum y \sum x^2 - \sum x \sum xy}{n(\sum x^2) - (\sum x)^2}$$

$$a = \frac{25 \times 120 - 20 \times 144}{4(120) - 400}$$

$$a = 1.5$$

Linear regression is given by:

$$y = a + bx$$

$$y = 1.5 + 0.95 x$$



# References

1. T. Mitchell, *Machine Learning*, McGraw-Hill, 1997
2. C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006
3. T. Jebarra, *Machine Learning: Discriminative and Generative*, Kluwer, 2004
4. R.O. Duda, P.E. Hart and D. Stork, *Pattern Classification*, 2<sup>nd</sup> Ed, Wiley 2002
5. C. Sutton and A. McCallum, *An Introduction to Conditional Random Fields for Relational Learning*
6. S. Shetty, H. Srinivasan and S. N. Srihari, *Handwritten Word Recognition using CRFs*, ICDAR 2007
7. S. Shetty, H. Srinivasan and S. N. Srihari, *Segmentation and Labeling of Documents using CRFs*, SPIE-DRR 2007