

---

# 6.867 Machine learning and neural networks

Tommi Jaakkola

MIT AI Lab

*tommi@ai.mit.edu*

Lecture 3: active learning, classification

---

# Topics

- Active learning and regression
  - sequential/batch
  - selection criteria
- Classification
  - Regression approach to classification

---

## Active learning: rules of the game

- Normal supervised learning:
  - (input,output) pairs are sampled from an *unknown joint* distribution  $P(x, y)$
- Active learning:
  - We can select the input examples, the corresponding outputs are sampled from an *unknown conditional* distribution  $P(y|x)$

---

# Active learning

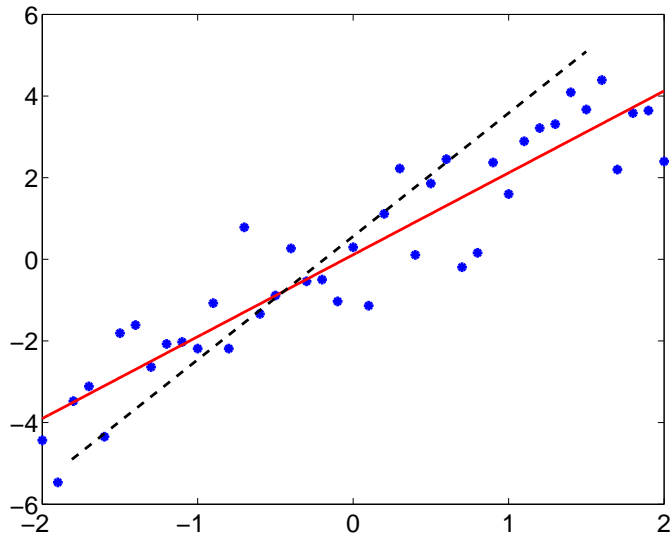
- Types of selection methods:
  1. **Batch selection:**

We select all the input examples prior to seeing any outputs
  2. **Sequential selection:**

We select each new input example on the basis of all the information so far
- We still need a specific selection criterion ...

## From previous lecture...

- Given a fixed set of input examples, the noise in the outputs generates variation in the estimated linear regression coefficients



Assumed “true” model:

$$\begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \dots & \dots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} w_0^* \\ w_1^* \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \dots \\ \epsilon_n \end{bmatrix}$$

$$\mathbf{y} = \mathbf{X}\mathbf{w}^* + \boldsymbol{\epsilon}$$

where  $\epsilon_i \sim N(0, \sigma^2)$ .

- Estimated linear coefficients:

$$\hat{\mathbf{w}} = \underbrace{\mathbf{w}^*}_{\text{true}} + \underbrace{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\epsilon}}_{\text{prediction from noise}}$$

- The estimated coefficients  $\hat{\mathbf{w}}$  are Gaussian random variables.

---

## From previous lecture... cont'd

$$\hat{\mathbf{w}} = \mathbf{w}^* + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon, \quad \epsilon \sim N(0, \sigma^2 I)$$

- We need to find the mean and the covariance of  $\hat{\mathbf{w}}$ :

$$\begin{aligned} E\{\hat{\mathbf{w}}\} &= \mathbf{w}^* \\ E\{(\hat{\mathbf{w}} - \mathbf{w}^*)(\hat{\mathbf{w}} - \mathbf{w}^*)^T\} &= E\left\{[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon] [(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon]^T\right\} \\ &= E\left\{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon \epsilon^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1}\right\} \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T E\{\epsilon \epsilon^T\} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\sigma^2 I) \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \\ &= \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \end{aligned}$$

- So, finally we get

$$\hat{\mathbf{w}} \sim N\left(\mathbf{w}^*, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}\right)$$

---

## Active learning: batch selection

We have to select the input examples prior to seeing any outputs

- We wish to find  $n$  inputs  $x_1, \dots, x_n$  (which determine the matrix  $\mathbf{X}$ ) so as to minimize some measure of randomness in the resulting coefficients  $\hat{\mathbf{w}}$

$$\hat{\mathbf{w}} \sim N \left( \mathbf{w}^*, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \right)$$

- What is the measure?

---

## Active learning: batch selection

We have to select the input examples prior to seeing any outputs

- We wish to find  $n$  inputs  $x_1, \dots, x_n$  (which determine the matrix  $\mathbf{X}$ ) so as to minimize some measure of randomness in the resulting coefficients  $\hat{\mathbf{w}}$

$$\hat{\mathbf{w}} \sim N \left( \mathbf{w}^*, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \right)$$

- What is the measure?

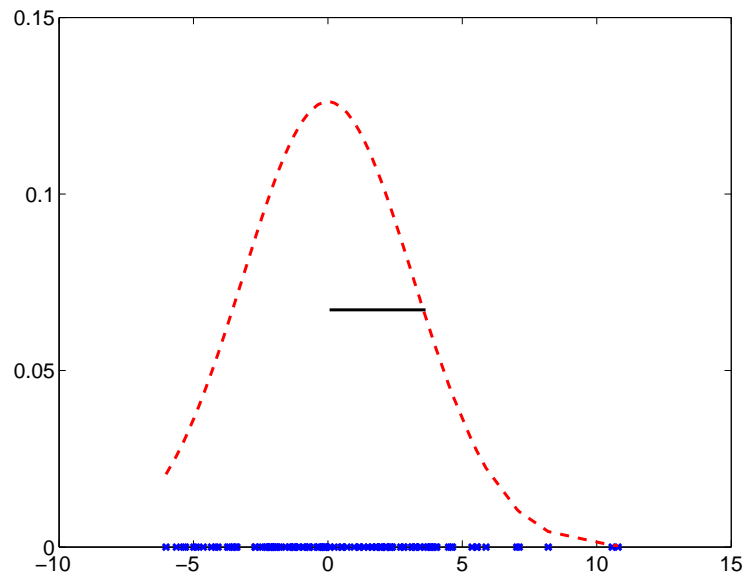
We find the points that minimize

$$\det \left[ (\mathbf{X}^T \mathbf{X})^{-1} \right]$$

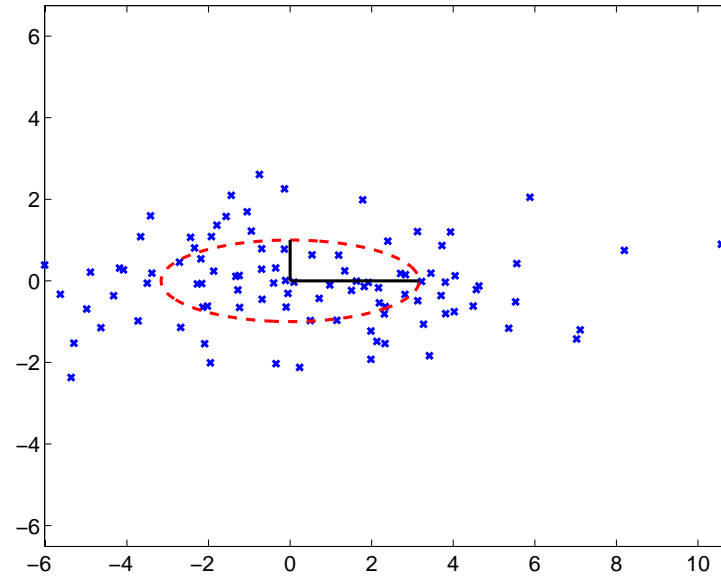


## Digression: “volume” of a Gaussian

- We can determine the “volume” of a Gaussian by looking at the covariance matrix



1-D (standard deviation)



2-D (product of stdv)

- More generally, “volume” is a function of the determinant of the covariance matrix

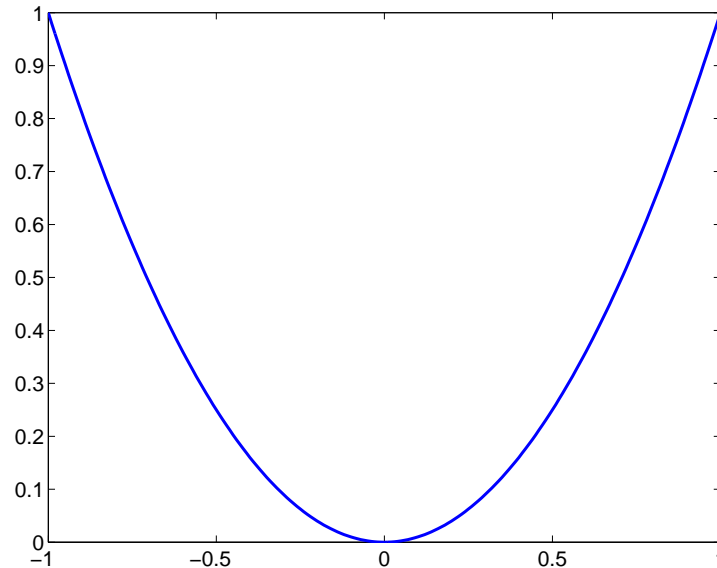
---

## Determinant criterion: example

- 1-d problem, 2nd order polynomial regression within  $x \in [-1, 1]$

$$f(x; \mathbf{w}) = w_0 + w_1x + w_2x^2$$

For  $n = 4$ , what points would we select?



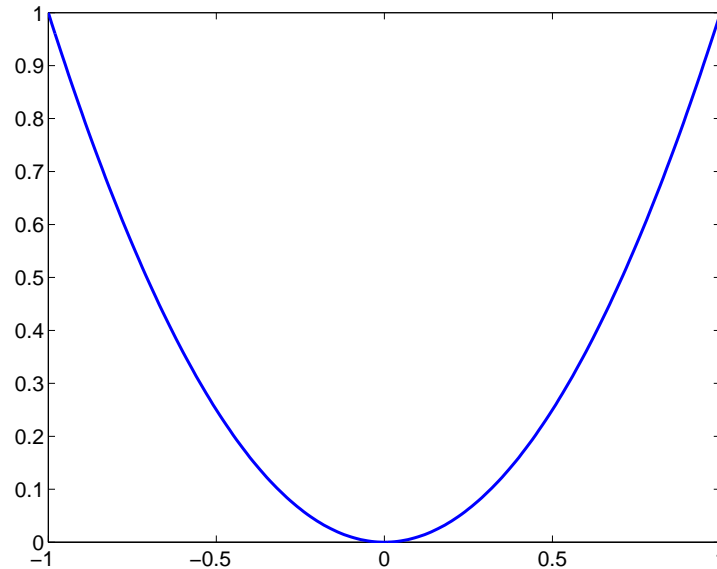
---

## Determinant criterion: example

- 1-d problem, 2nd order polynomial regression within  $x \in [-1, 1]$

$$f(x; \mathbf{w}) = w_0 + w_1x + w_2x^2$$

For  $n = 4$ , what points would we select?



$$x_1 = -1, x_2 = 0, x_3 = 0, x_4 = 1$$

---

## Active learning: sequential selection

We can select the next input example on the basis of all the inputs and outputs already observed

---

## Active learning: sequential selection

We can select the next input example on the basis of all the inputs and outputs already observed

- We can also select the input points to reduce the variance in our *predictions*

$$\hat{y}(x) = \hat{w}_0 + \hat{w}_1 x = \begin{bmatrix} 1 \\ x \end{bmatrix}^T \begin{bmatrix} \hat{w}_0 \\ \hat{w}_1 \end{bmatrix}$$

The variance in the prediction at  $x$  is

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \end{bmatrix} \quad \text{Var} \{ \hat{y}(x) \} = \sigma^2 \begin{bmatrix} 1 \\ x \end{bmatrix}^T (\mathbf{X}^T \mathbf{X})^{-1} \begin{bmatrix} 1 \\ x \end{bmatrix}$$

- the noise variance  $\sigma^2$  only affects the overall scale
- the variance is a function of previously chosen inputs, not outputs!

---

## Active learning: sequential selection

We can select the next input example on the basis of all the inputs and outputs already observed

- We can also select the input points to reduce the variance in our *predictions*

$$\hat{y}(x) = \hat{w}_0 + \hat{w}_1 x = \begin{bmatrix} 1 \\ x \end{bmatrix}^T \begin{bmatrix} \hat{w}_0 \\ \hat{w}_1 \end{bmatrix}$$

The variance in the prediction at  $x$  is

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \end{bmatrix} \quad \text{Var} \{ \hat{y}(x) \} = \sigma^2 \begin{bmatrix} 1 \\ x \end{bmatrix}^T (\mathbf{X}^T \mathbf{X})^{-1} \begin{bmatrix} 1 \\ x \end{bmatrix}$$

- the noise variance  $\sigma^2$  only affects the overall scale
- the variance is a function of previously chosen inputs, not outputs!

- The selection criterion:

$$x^{new} = \arg \max_x \{ \text{Var} \{ \hat{y}(x) \} \}$$

---

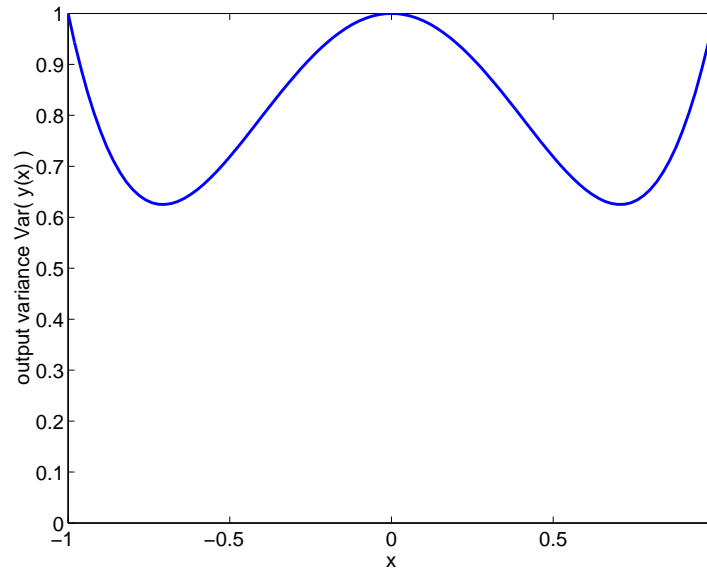
## Sequential selection: example

- 1-d problem, 2nd order polynomial regression within  $x \in [-1, 1]$

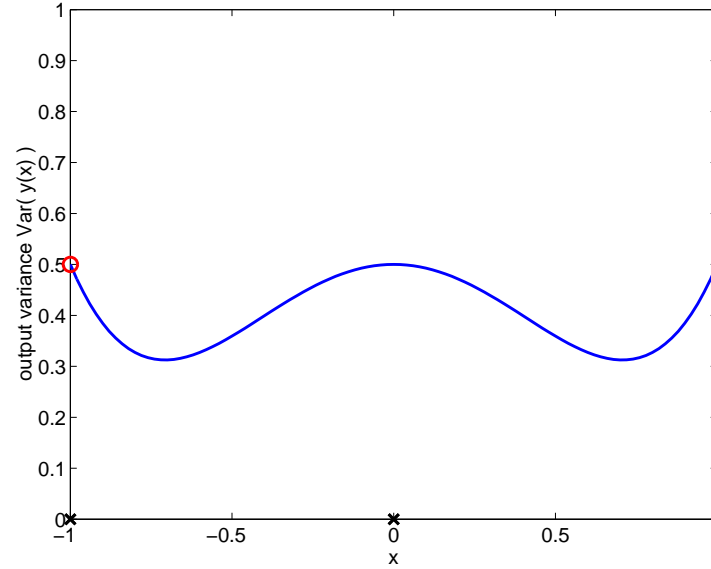
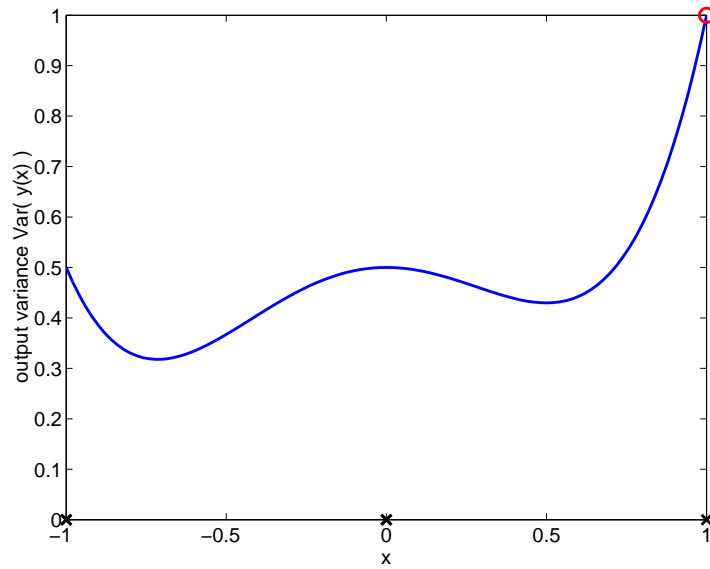
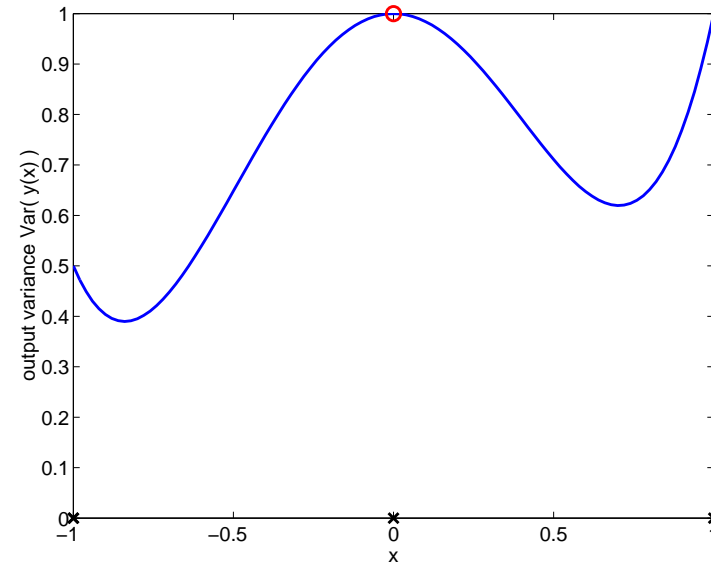
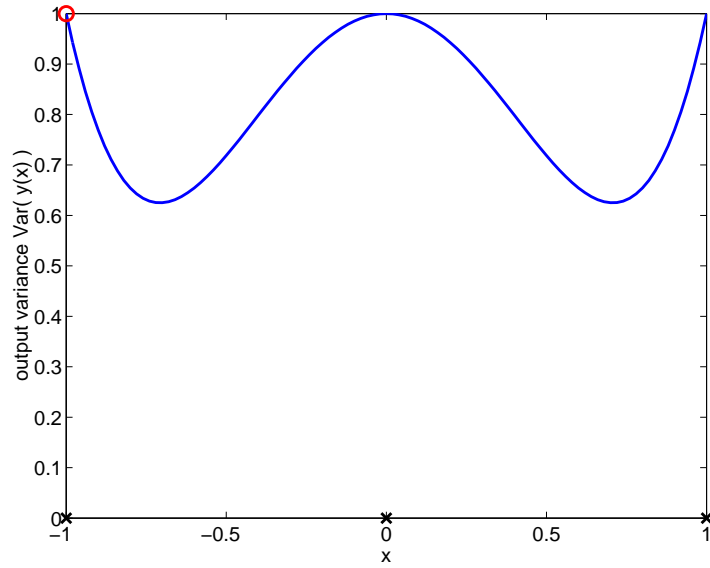
$$\hat{y}(x) = \hat{w}_0 + \hat{w}_1 x + \hat{w}_2 x^2$$

A priori selected inputs  $x_1 = -1, x_2 = 0, x_3 = 1$ .

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \dots & \dots & \dots \end{bmatrix} \quad \text{Var} \{ \hat{y}(x) \} = \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix}^T (\mathbf{X}^T \mathbf{X})^{-1} \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix}$$



# Example cont'd





---

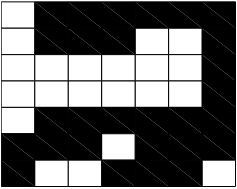
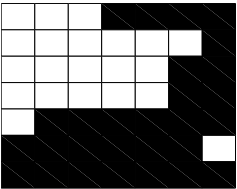
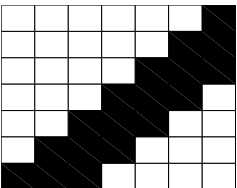
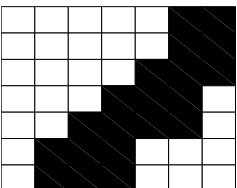
# Topics

- Classification
  - Regression approach to classification

---

# Classification

Example: digit recognition (8x8 binary digits)

binary digit	actual label	target label in learning
	"2"	1
	"2"	1
	"1"	0
	"1"	0
...	...	

---

## Classification via regression

- We ignore the fact that the output is binary (e.g., 0/1) rather than a continuous variable

Given a linear regression function

$$f(\mathbf{x}; \mathbf{w}) = w_0 + w_1x_1 + \dots + w_dx_d$$

we minimize the squared difference between the predicted output (continuous) and the observed label (binary):

$$J_n(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^n (y_i - f(\mathbf{x}_i; \mathbf{w}))^2$$

- How do we classify any new example  $\mathbf{x}$ ?

---

## Classification via regression cont'd

$$f(\mathbf{x}; \mathbf{w}) = w_0 + w_1x_1 + \dots + w_dx_d$$

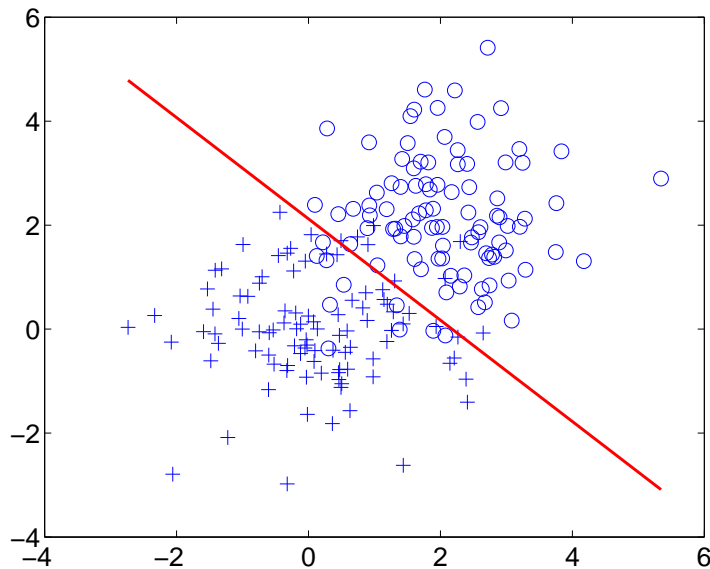
Any new (test) example  $\mathbf{x}$  can be classified according to

label = 1 if  $f(\mathbf{x}; \mathbf{w}) > 0.5$ , and label = 0 otherwise

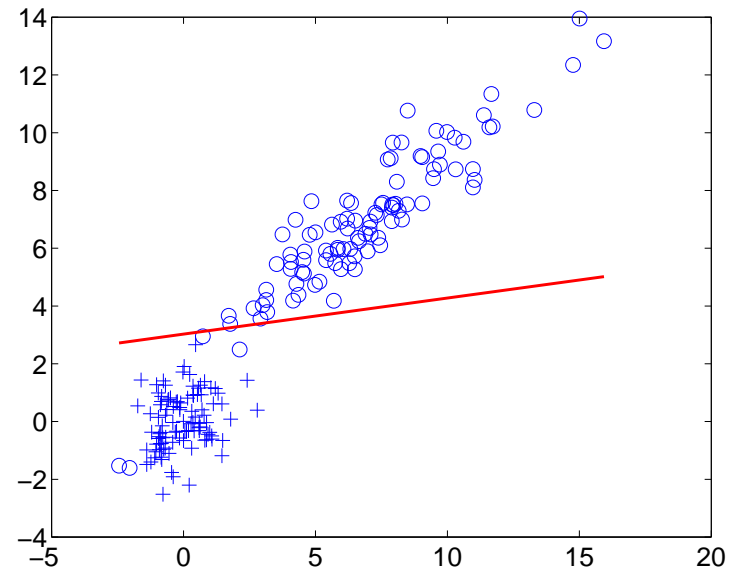
where  $f(\mathbf{x}; \mathbf{w}) = 0.5$  defines the **decision boundary**.

# Classification via regression cont'd

- This is not optimal... why not?



sometimes good



sometimes bad