University of Amsterdam
Informatics Institute

AMLAB
Amsterdam Machine Learning Lab

Machine Learning 1

Lecture 6.4 - Supervised Learning Classification - Discriminative Models - Least Squares Regression

Erik Bekkers

(Bishop 4.1.3)

Image credit: Kirillm | Getty Images

Slide credits: Patrick Forré and Rianne van den Berg

Least Squares for Classification (K)

- Each class C_k has its own linear model:

$$y_k(\mathbf{x}) = \mathbf{w}_k^T \mathbf{x} + w_{k0}$$

- Shorter notation: $\mathbf{y}(\mathbf{x}) = \widetilde{\mathbf{W}}^T \tilde{\mathbf{x}}$
- Matrix $\widetilde{\mathbf{W}}$: column *k* contains $\tilde{\mathbf{w}}_k = (w_{k0}, \underline{w})^T \in \mathbb{R}^M$
- Vector $\tilde{\mathbf{x}} = (1, \underline{x})^T \in \mathbb{R}^M$
- Vector $\mathbf{y}(\mathbf{x}) = \begin{pmatrix} y_1(\underline{x}) \\ y_2(\underline{x}) \\ \vdots \\ y_K(\underline{x}) \end{pmatrix} = \tilde{W}^T \tilde{x} \in \mathbb{R}^K$
- Assign $\mathbf{x}$ to class C_k if

$$k = \underset{j}{\operatorname{argmax}}\, y_j(\underline{x})$$

Least Squares for Classification (II)

- Data set: N x (D+1) data matrix, N x K target matrix

$t_n = (0,0,1,0,0)^T$

$$\widetilde{X} = \begin{pmatrix} -\tilde{x}_1^T- \\ \vdots \\ -\tilde{x}_N^T- \end{pmatrix} \qquad T = \begin{pmatrix} -t_1^T- \\ \vdots \\ -t_N^T- \end{pmatrix}$$

$$(\tilde{X}\tilde{W} - T)_{nk} = \sum_m X_{nm} W_{mk} - T_{nk}$$

- Use sum-of-squares regression error function

$$E_D(\widetilde{\mathbf{W}}) = \frac{1}{2}\mathrm{Tr}\left[\underbrace{\overbrace{(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})^T}^{K \times N}\overbrace{(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})}^{N \times K}}_{K \times K}\right]$$

$$= \frac{1}{2}\sum_{k=1}^{K}\sum_{n=1}^{N}\left(\sum_{m=1}^{M}\tilde{X}_{nm}\tilde{W}_{mk} - T_{nk}\right)^2$$

- Minimize $E_D(\widetilde{\mathbf{W}})$ as a function of $\widetilde{\mathbf{W}}$:

$$\frac{\partial E_D(\tilde{W})}{\partial \tilde{W}} = 0$$

- Solution: $\widetilde{\mathbf{W}}_{\mathrm{LS}} = \left(\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}\right)^{-1}\widetilde{\mathbf{X}}^T\mathbf{T} = \widetilde{\mathbf{X}}^\dagger\mathbf{T}$

- Discriminant function: $\mathbf{y}_{\mathrm{LS}}(\mathbf{x}) = \tilde{W}_{LS}^T \tilde{x}$

Least Squares for Classification: Problems

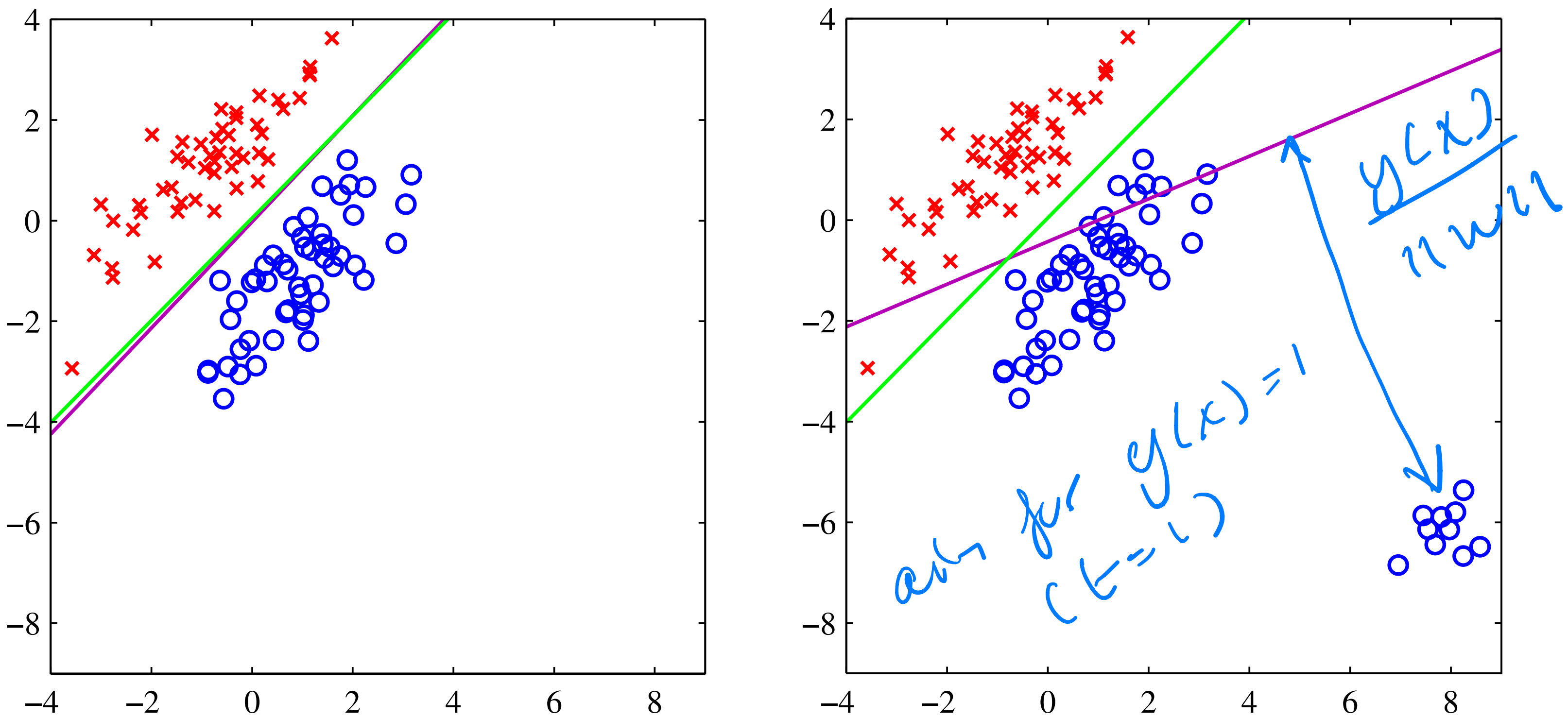

Figure: least squares is very sensitive to outliers (Bishop 4.4)

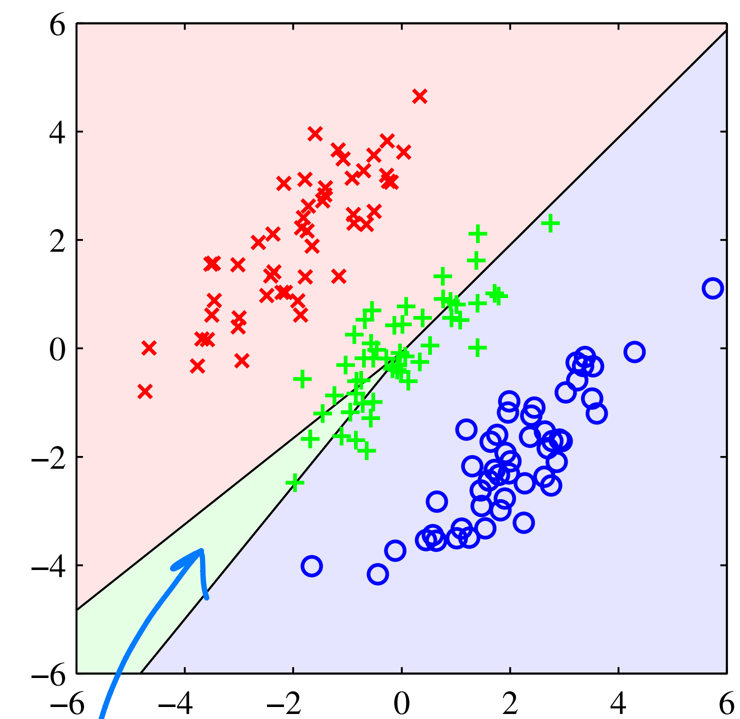

Figure: masking for least squares for K>2 (Bishop 4.5)

masking

1. The decision boundaries are very sensitive to outliers

2. For K>2 some decision regions can become very small or are even completely ignored

3. The components of $\mathbf{y}_{\mathrm{LS}}(\mathbf{x})$ are not real probabilities!
   - $y_k(\mathbf{x})$
   - if $\sum_{k=1}^{K} t_k = 1$ $\rightarrow$ $\sum_k y_k = 1$