Classification Lab

Luca Morreale

Linear classifier

$$z(\mathbf{x}) = \mathbf{w}^\top(\mathbf{x}|1) = \sum_i w_i \cdot x_i$$

$$z(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$$

Taken from Tobias Ritschel Slides

Linear classifier

$$z(\mathbf{x}) = \mathbf{w}^{\mathsf{T}}\mathbf{x}$$

Here X is assumed to be FxN where:

N is the number of samples

F is the number of features

In practice that may not be the case!

A simple **solution** is to **transpose** X

$$\mathrm{x} = X^T$$

Optimization

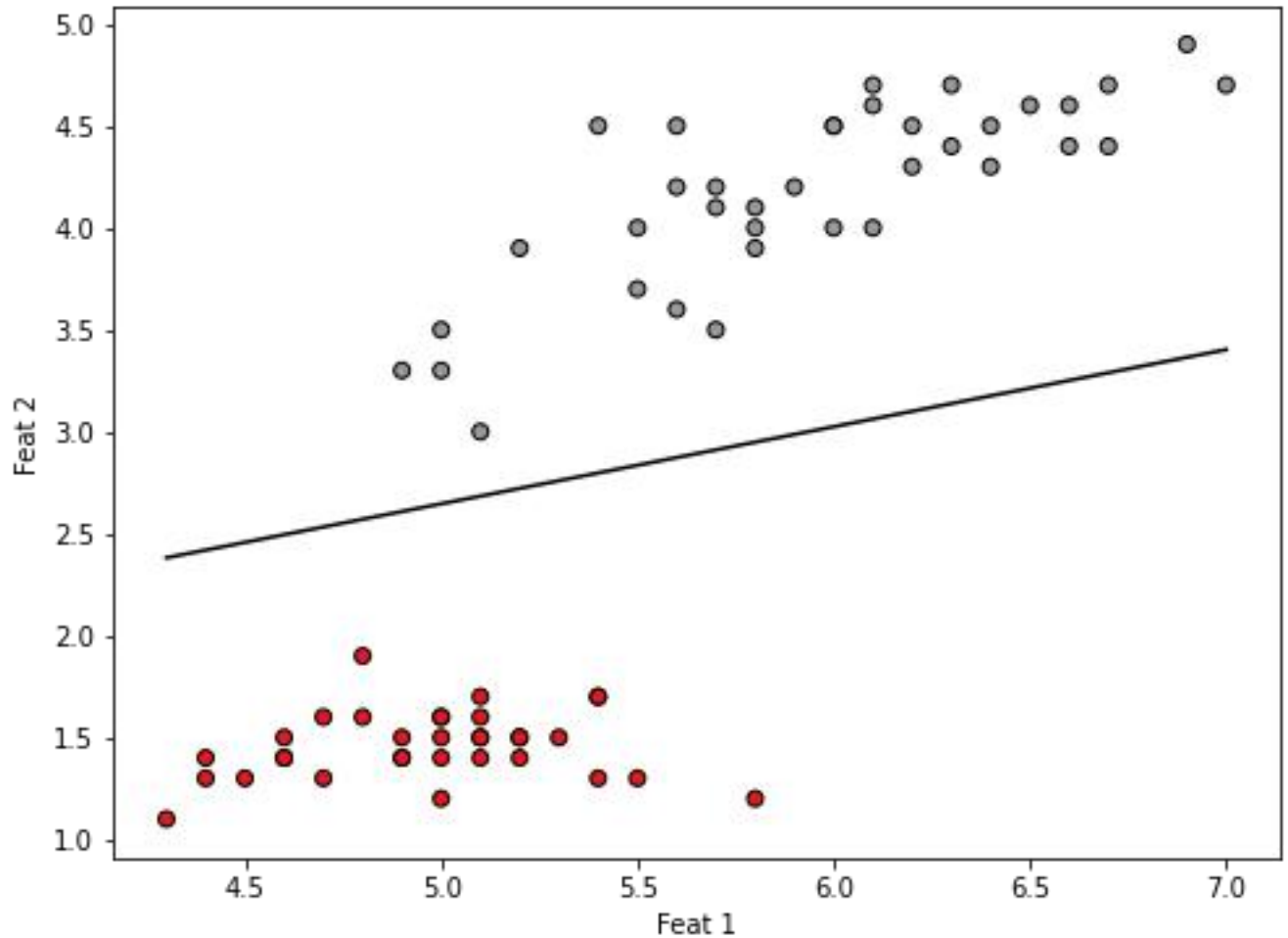
- Finding w is a least squares (LS) problem

$$\text{argmin}_{\mathbf{w}} \sum_i ||\mathbf{z}_i - \mathbf{w}^\mathsf{T}\mathbf{x}_i||_2$$

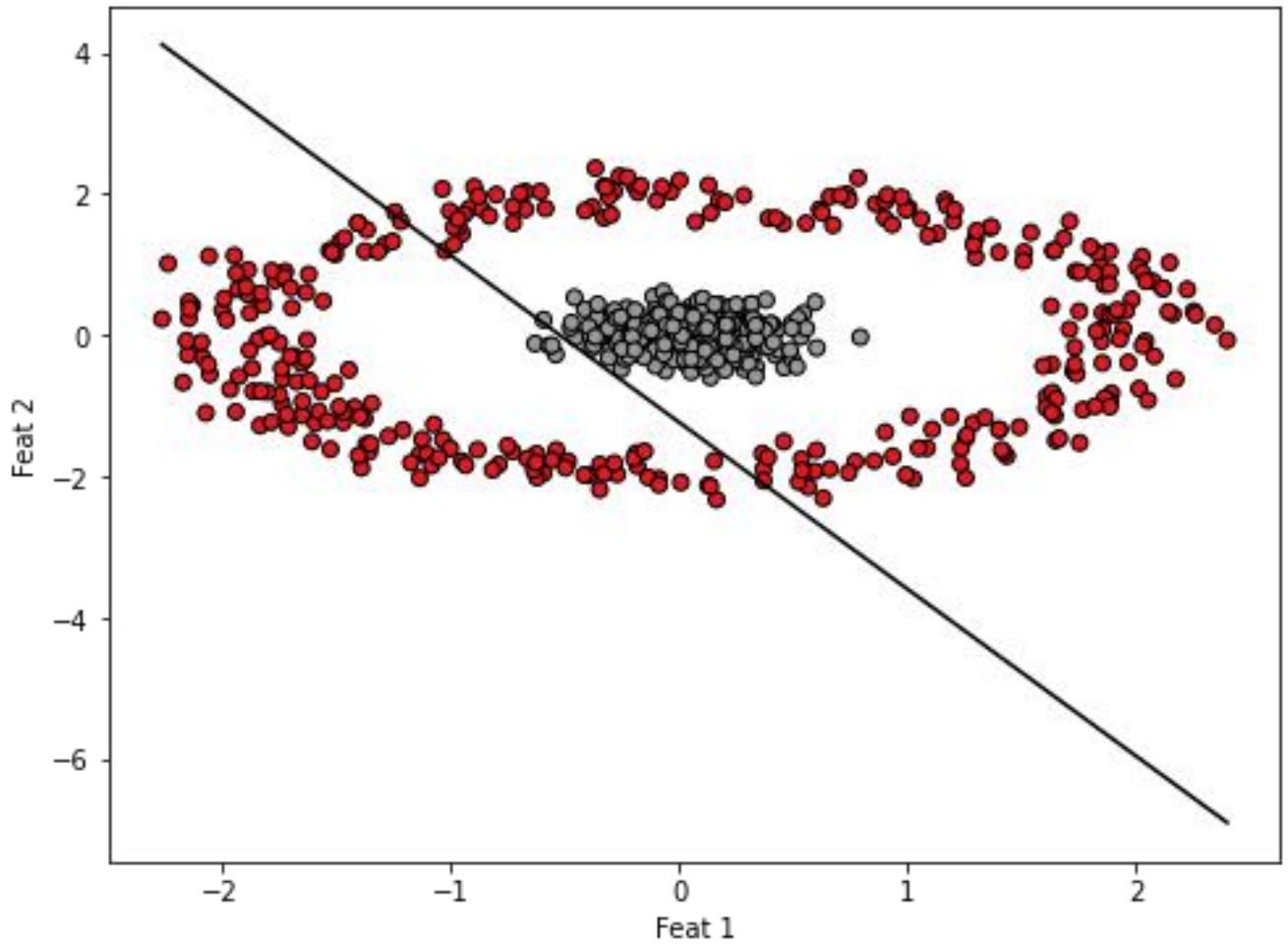
$$\mathbf{w}_{\text{LS}} = (\mathsf{X}\mathsf{X}^\mathsf{T})^{-1}\mathsf{X}\mathbf{y}$$

Taken from Tobias Ritschel Slides

Decision boundary



Decision boundary



Logistic Regression

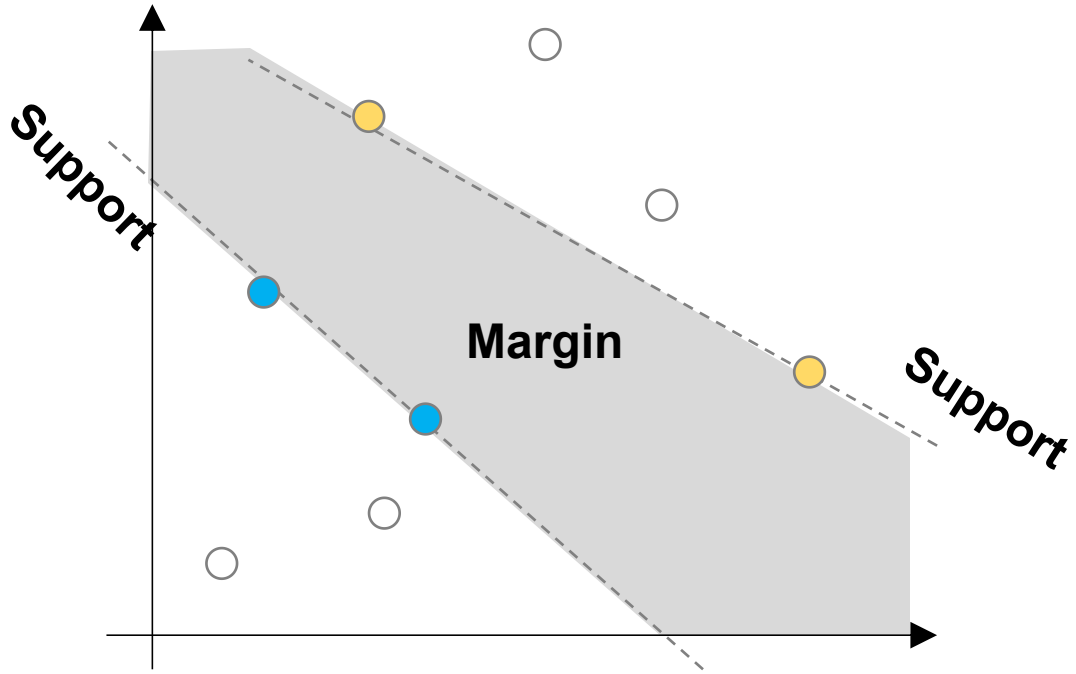
- Finding **w** now required non-linear least squares (e.g. Gradient descent)
- Has analytic gradients

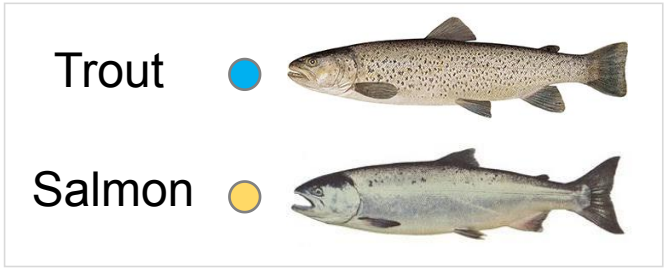
$$z = \frac{1}{1 - \exp(-\mathbf{x}^\mathsf{T}\mathbf{w})}$$

Taken from Tobias Ritschel Slides

Support and Margin

- **“Support vectors”** is the exemplar subset that is hard to classify
- **“Margin”** is the “street” between them

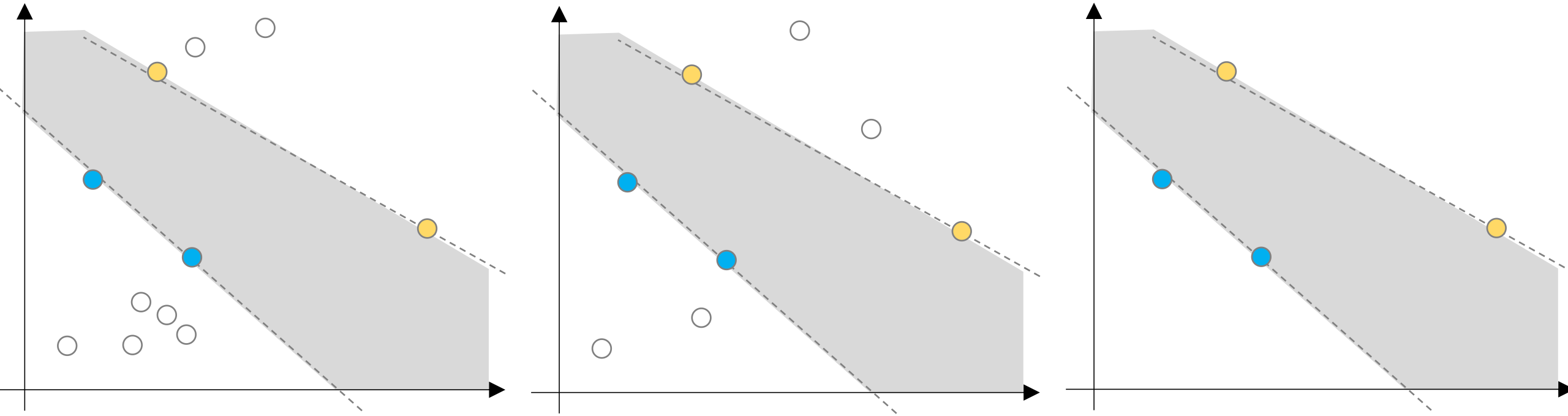




Taken from Tobias Ritschel Slides

Support and Margin

- Only the support vectors matter



Taken from Tobias Ritschel Slides

Kernel Trick, Definitions 2

- Remember the dual form of the SVM

$$z(\mathbf{x}) = \sum_i \alpha_i z_i \mathbf{x}_i^\mathsf{T} \mathbf{x}$$

- We can now use

$$z(\mathbf{x}) = \sum_i \alpha_i z_i \left\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}) \right\rangle = \sum_i \alpha_i z_i \kappa(\mathbf{x}_i, \mathbf{x}_i)$$

Taken from Tobias Ritschel Slides

Other Kernels

- Any function $\kappa(\mathbf{x}_1, \mathbf{x}_2)$ is a kernel, as long its kernel matrix is **symmetric positive semi-definite**
- Other kernels:
  - Polynomial
  - Radial basis function
  - Sigmoid

$$\kappa(\mathbf{x}_1, \mathbf{x}_2)_{\text{Poly}} = (1 + \mathbf{x}_1^\top \mathbf{x}_2)^p$$

$$\kappa(\mathbf{x}_1, \mathbf{x}_2)_{\text{Radial}} = \exp\left(\frac{||\mathbf{x}_1 - \mathbf{x}_2||^2}{2\sigma}\right)$$

$$\kappa(\mathbf{x}_1, \mathbf{x}_2)_{\text{Sigmoid}} = \tanh(\beta_0 + \beta_1 \mathbf{x}_1^\top \mathbf{x}_2)$$

Taken from Tobias Ritschel Slides

Classification Metrics

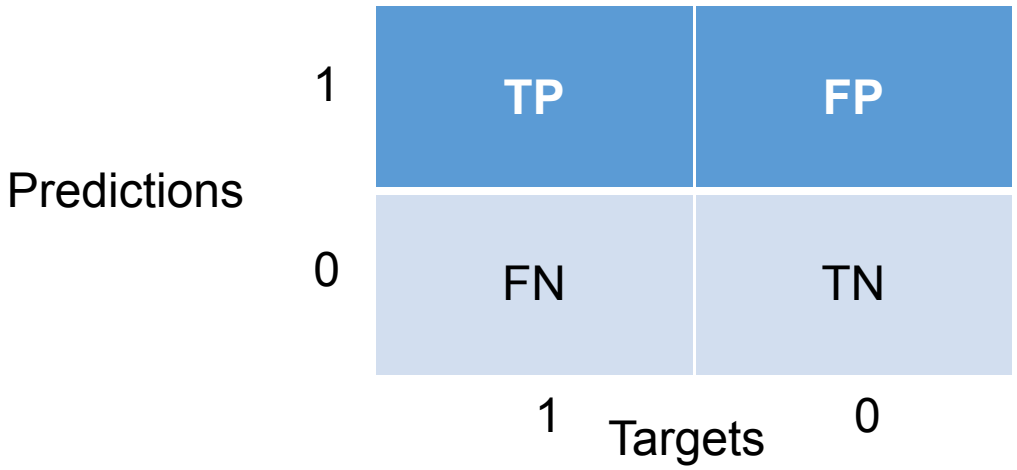
- Once we have classifier we need to **numerically** evaluate it.

- How?
  - Confusion Matrix
  - Accuracy
  - Precision

Confusion Matrix

Give a complete overview of model performances

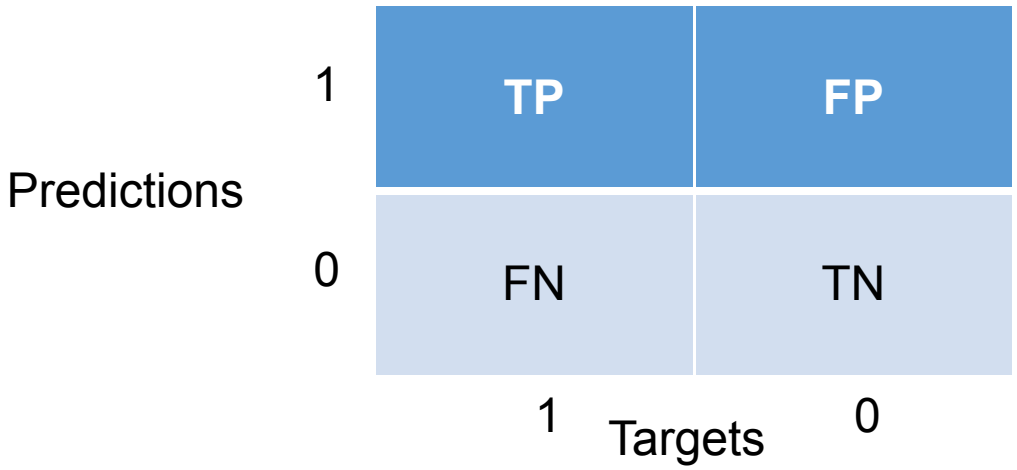
- How many sample **positive** are predicted **positive**? **TP**
- How many sample **negative** are predicted **negative**? **TN**
- How many sample **positive** are predicted **negative**? **FP**
- How many sample **negative** are predicted **positive**? **FN**



Confusion Matrix (few notes)

- Entries order can change!
- Axis can be swapped!
  - **Always specify axis and values**

- Matrix show works for binary, can be extended to multi-class.



Accuracy

How many **correct predictions** do we have?

- The key is **correct**
  - sample positive predicted positive (TP)
  - sample negative predicted negative (TN)
  - compare to total number of samples

$$A = \frac{TP + TN}{N}$$

- What happens when the dataset is unbalanced? (#P > #N)

Precision

How many **positive predictions** do we have?

- The key is **positive**
  - sample positive predicted positive (TP)
  - compare to positive predicted samples (TP+FP)

$$P = \frac{TP}{TP + FP}$$

- What happens when the dataset is unbalanced? (#P > #N)

There are many more measure

- There are plenty measure to quantify the quality of your model
  - Recall
  - F1 score
  - ROC curve
  - ....
- We are not diving into all of them
- Note some of them are suffer in case unbalanced dataset