Today: Start classification

Recap regression via least squares methods

Find $f: \mathbb{R}^d \to \mathbb{R}$ so that

$$f(\vec{x}_i) \approx y_i \quad \text{for} \quad (\vec{x}_i, y_i)_{i=1}^{n}$$

training data

$$\min_f \frac{1}{n} \sum_{i=1}^{n} (f(\vec{x}_i) - y_i)^2$$

Mean squared error $\longrightarrow$ least squares estimator

Models for f

intercept
offset

coefficient

- linear model $f(\vec{x}) = \beta_0 + \beta_1 x_1 + \dots + \beta_d x_d$

$$= [1, x_1, \dots, x_d] \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{bmatrix}$$

$$= \vec{x}^T \vec{\beta}$$

parametric model $\vec{\beta}$

= fixed dimensions

- fixed features + linear model

basis

$$f(\vec{x}) = \vec{\Phi}(\vec{x})^T \vec{\beta} = \sum_{i=1}^{f} \phi_i(\vec{x}) \beta_i$$

ex/ degree-2 polynomials

basis fxns

feature vector

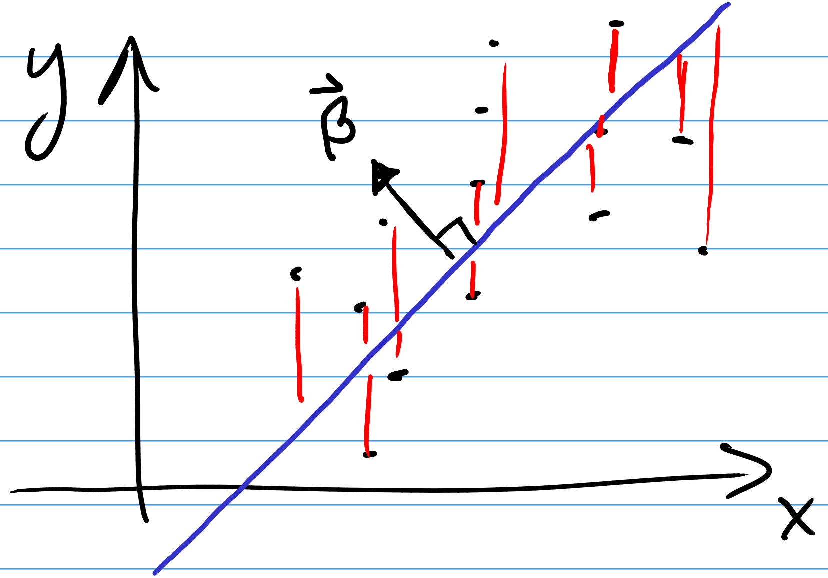
$$\vec{\Phi}(\vec{x}) = [1, x_1, x_2, x_1^2, x_1 x_2, x_2^2]^T \in \mathbb{R}^6$$

featurization

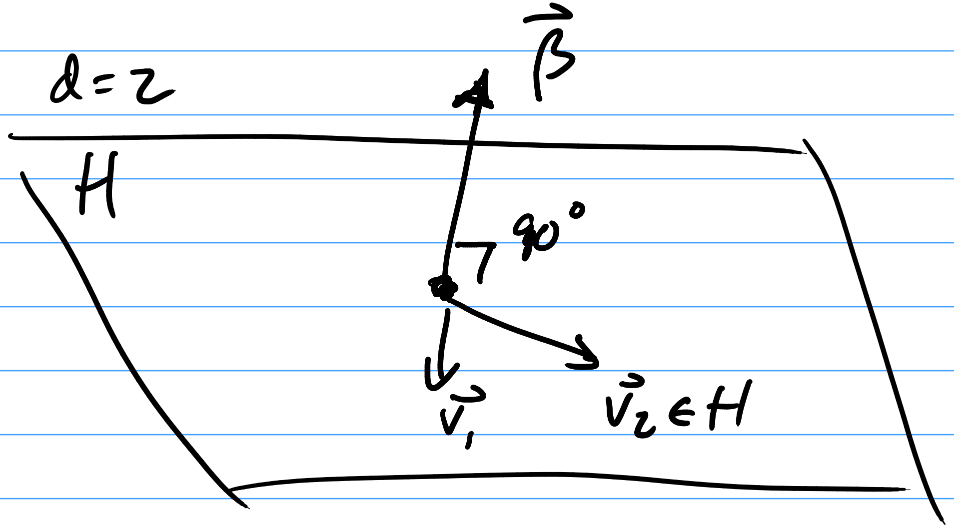
$$\vec{x} \in \mathbb{R}^2, \quad d = 2, \quad \vec{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Linear predictors fit hyperplane to $(\vec{x}_i, y)$

<u>$d=1$</u>



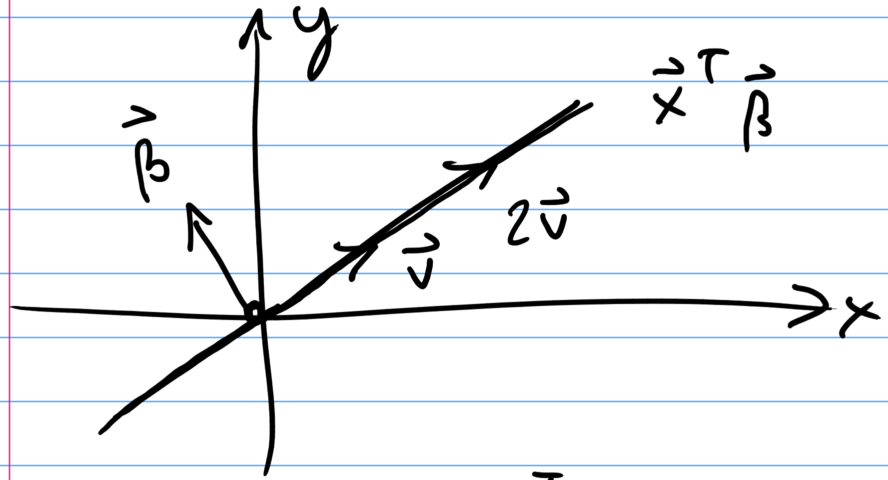
in $d=1$ dims, hyperplane is a line



$\vec{\beta}$ is orthogonal to the hyperplane

$\vec{\beta}^T \vec{v} = 0$ for any $\vec{v} \in H$

If $\beta_0 = 0$ then $\vec{\beta}$



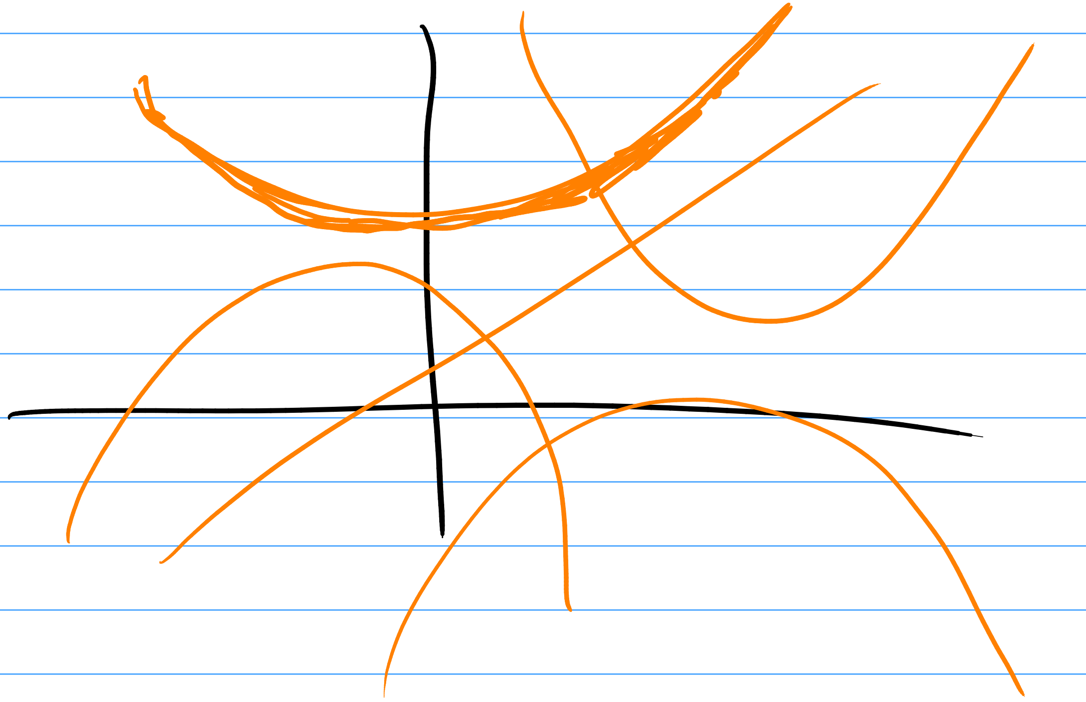
orthogonal : $\vec{\beta}^T \vec{v} = 0$

orthonormal : $\|\vec{\beta}\| = 1$

$\|\vec{v}\| = 1$

$\phi_i(\vec{x})$

$3 \cdot 1 - \frac{1}{2} \cdot x + 1 x^2 = 3 + x^2 - \frac{1}{2} x$



basis of nonlinear features

Estimators:

- Ordinary least squares $\quad \min_{\vec{\beta}} \|\underset{\text{features}}{X}\vec{\beta} - \underset{\text{labels}}{\vec{y}}\|^2$

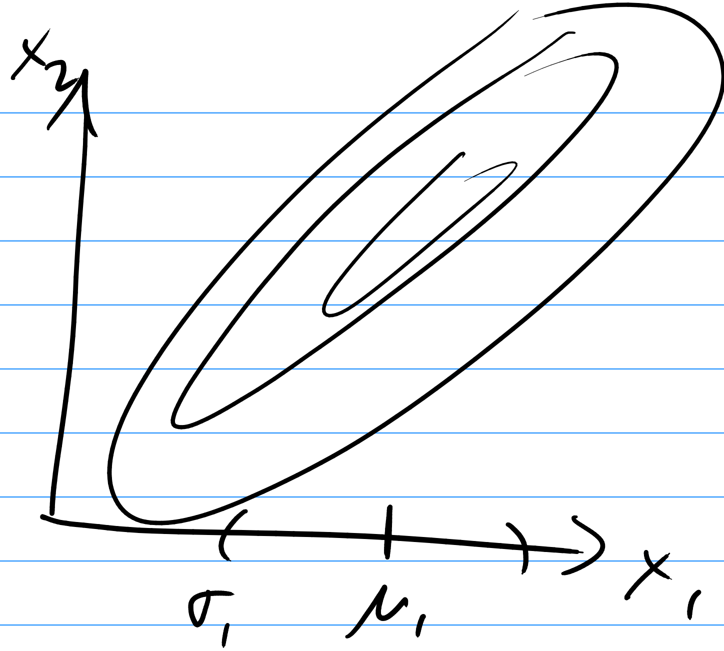
- Ridge regression $\quad \dots + \underbrace{\lambda \|\vec{\beta}\|^2}_{\text{regularization penalty}}$

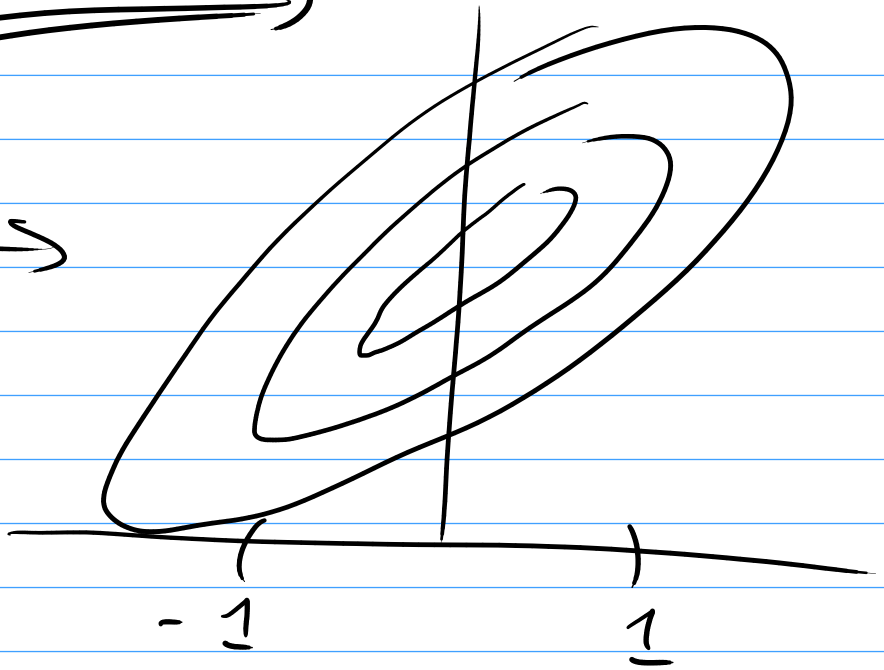
  controlling variance
  noise in y's
  collinear features
  - small singular values of X

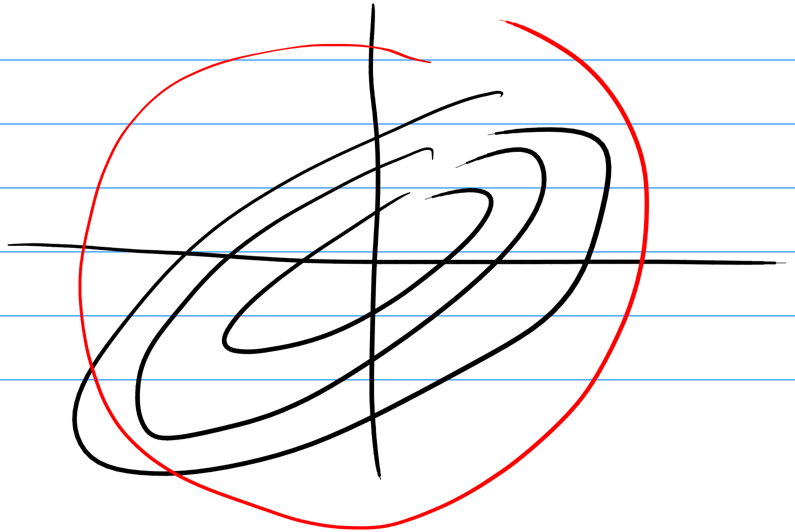
- Lasso, giving sparsity $\quad \dots + \lambda \|\vec{\beta}\|_1$

$$\vec{\beta}_{lasso} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

Preprocessing







Standardization

vs. whitening
makes singular values equal

Turns out that least squares $\min_{\hat{y}} (\hat{y} - y)$

prediction

Best possible predictor

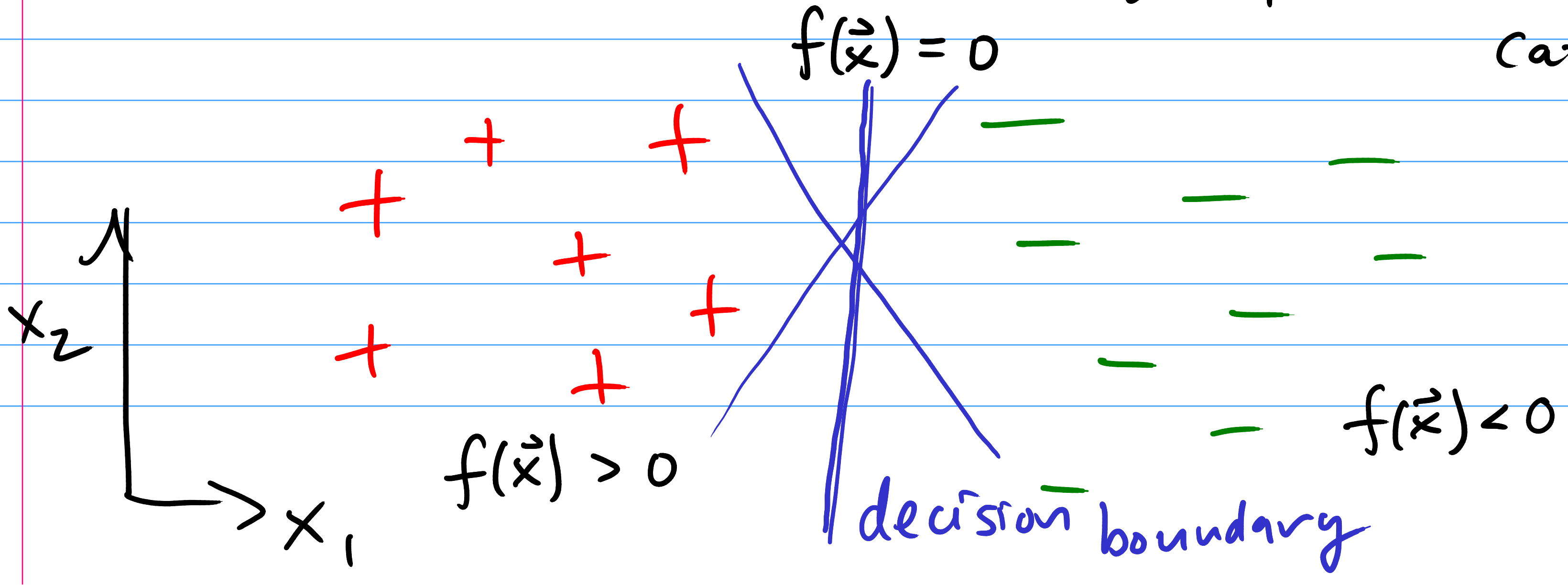
$$f(\vec{x}) = \mathbb{E}[Y \mid \vec{x}]$$

Bayes predictor

Classification

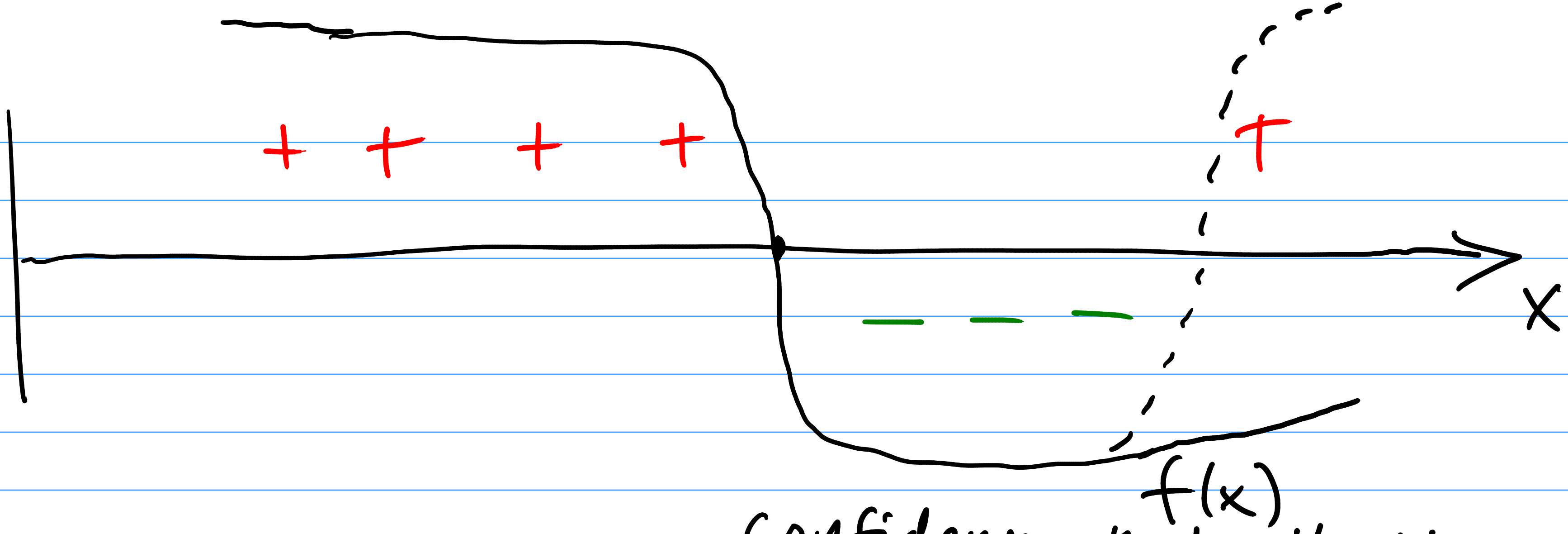
$$y = \begin{cases} +1 & \text{"true"} \\ -1 & \text{"false"} \end{cases}$$

categorical



Want f to model

$$\Pr[Y = +1 \mid \vec{x}]$$



	confidence that $y=+1$
$f(x_1) = 10$	high
$f(x_2) = 1$	med
$f(x_3) = 0.1$	low

5

n_{test} vs. n_{train}

$$\frac{1}{n}\sum_{i=1}^{n}\left(f(\vec{x}_i) - y_i\right)^2$$

$$\frac{1}{n}\left\|X\vec{\beta} - \vec{y}\right\|^2$$

arrays

np.linalg.mean((ypred - y)**2)

MSE mean square error

2.3-4 very similar

4.5

$Y = [2 \times 61 \quad] =$