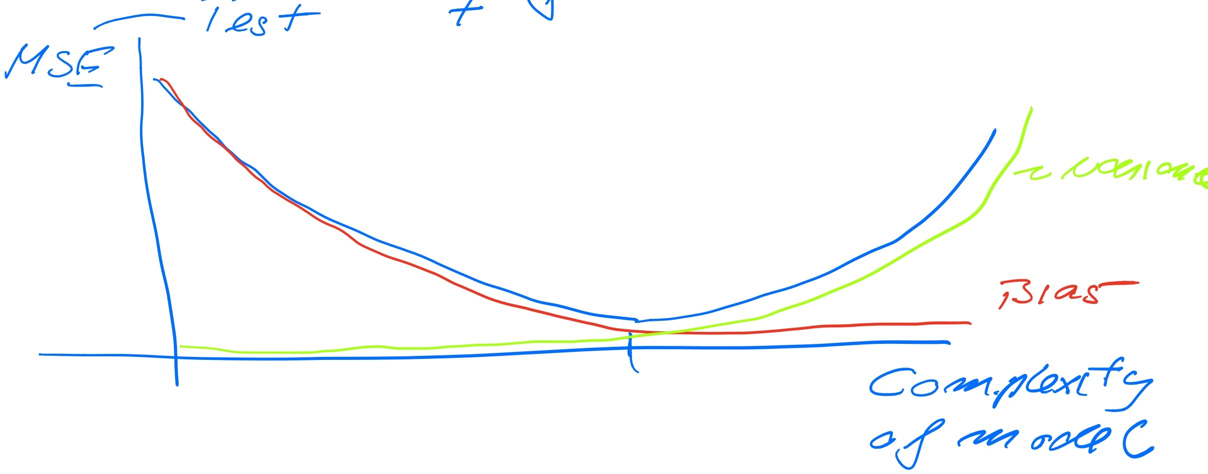
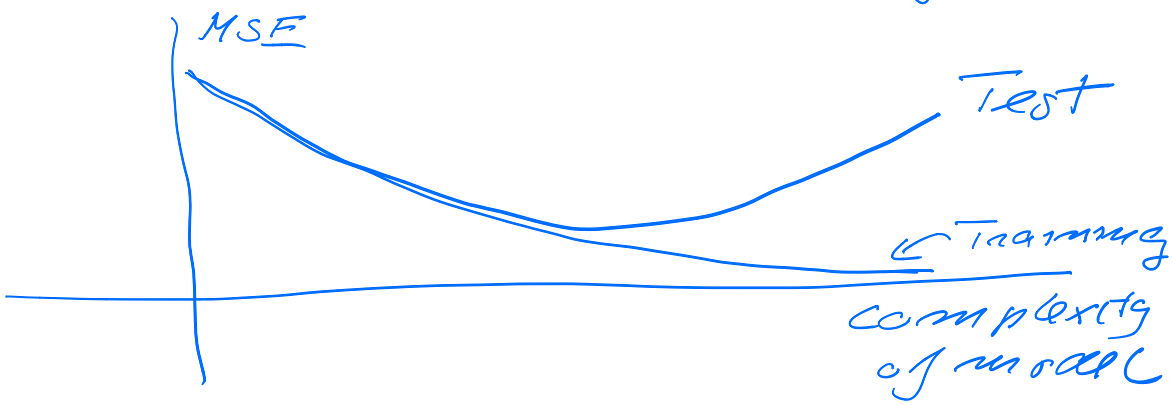
Bias-variance tradeoff

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 = \mathbb{E}\left[(y-\tilde{y})^2\right]$$

$$= \underbrace{\mathbb{E}\left[(y-\mathbb{E}[\tilde{y}])^2\right]}_{\text{Bias}} + \underbrace{\mathbb{E}\left[(\tilde{y}-\mathbb{E}[\tilde{y}])^2\right]}_{\text{variance}} + \sigma^2$$





- ordinary least squares (OLS)
  - Design matrix $X \in \mathbb{R}^{n \times p}$
  - output $y \in \mathbb{R}^n$
  - Estimator $\beta \in \mathbb{R}^p$

$$\beta = (X^T X)^{-1} X^T y$$

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \left\{ (y - X\beta)^T (y - X\beta) \right\}$$

OLS

$$= \frac{1}{n} \| (y - X\beta) \|_2^2 = C(\beta)$$

$$\|x\|_2 = \sqrt{\sum_i x_i^2}$$

Ridge

$$\min_{\beta \in \mathbb{R}^p} = \boxed{\frac{1}{n} \| y - X\beta \|_2^2 + \underline{\lambda \beta^T \beta}}$$

$$\lambda \|\beta\|_2^2$$

For $\lambda > 0$ $\quad MSE(\lambda) < MSE(OLS)$

$MSE(Ridge)$

OLS

$$\frac{\partial C(\beta)}{\partial \beta} \Rightarrow \beta = \underline{(X^T X)}^{-1} X^T y$$

$$\frac{\partial C(\beta)}{\partial \beta} = 0 = (X^T (y - X\beta)) \times (-2)$$

Ridge

$$\frac{\partial C(\beta, \lambda)}{\partial \beta} = 0 = X^T (y - X\beta) - \lambda \beta$$

$$X^T X\beta + \lambda\beta = X^T y \Rightarrow$$

$$\beta = (X^T X + \lambda I)^{-1} X^T y$$

with finite $\lambda > 0$, no problem with divergencies

the matrix $X^T X + \lambda I$

$$X^T X \in \mathbb{R}^{p\times p} \quad I \in \mathbb{R}^{p\times p}$$

Lasso optimization;

$$C(\beta,\lambda) = \frac{1}{n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

$$\|x\|_1 = \sum_i |x_i|$$

---

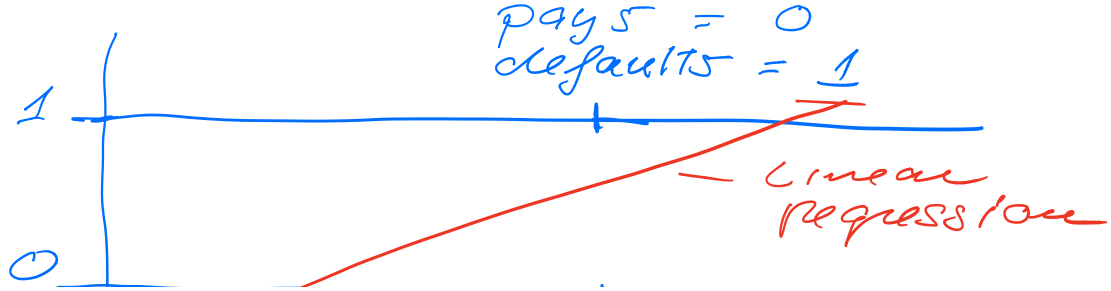
classification problem

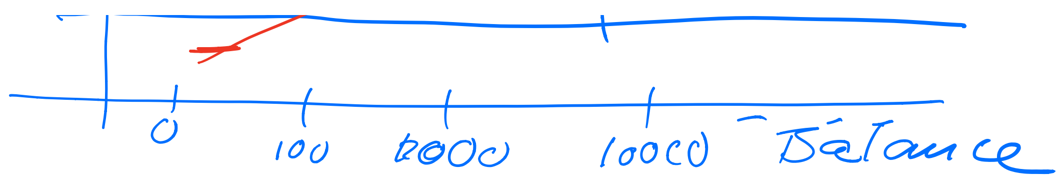
(Logistic regression)

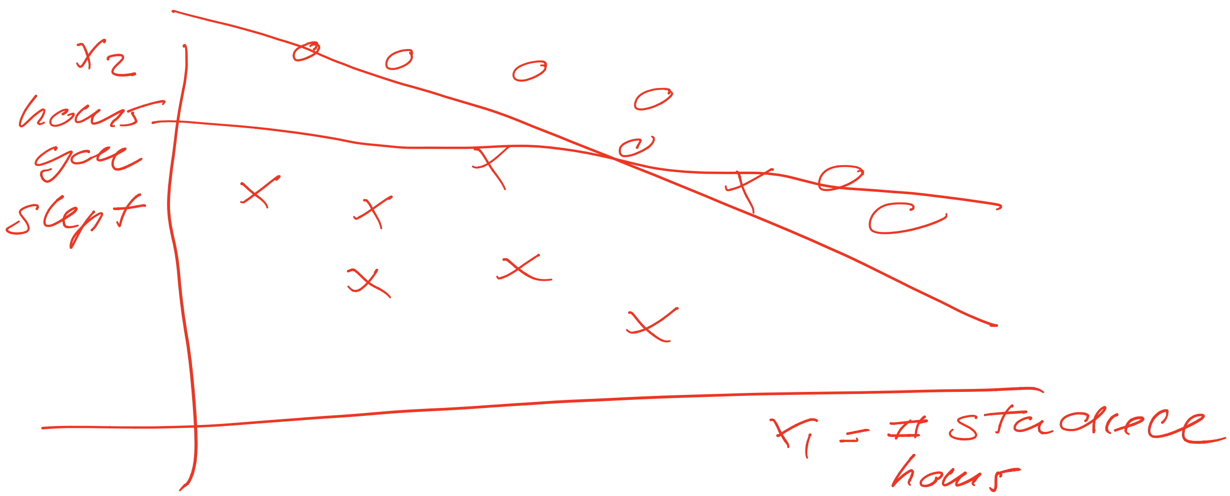
Example credit card data

pays = 0

defaults = 1







Logistic Regression

$$p(t) = \frac{1}{1+e^{-t}} = \frac{e^t}{1+e^t}$$

$$1 - p(t) = p(-t)$$

$$t_i = \beta_0 + \beta_1 x_i \qquad \text{Model}$$

Probability $\qquad \beta_0 \; \beta_1$

$$P(y_i = 1 \mid x_i \beta) = \frac{e^{t_i}}{1+e^{t_i}}$$

$$P(y_i = 0 \mid x_i \beta) = 1 - P(y_i = 1 \mid x_i \beta)$$

Maximum likelihood approx
(iid)

$$D = \{(y_0 x_0), (y_1 x_1) \ldots (y_{n-1} x_{n-1})\}$$

Probability

$$P(D|\beta) =$$

$$\prod_{i=0}^{n-1} \left(P(y_i=1|x_i\beta)\right)^{y_i} \left(1-P(y_i=1|x_i\beta)\right)^{1-y_i}$$

How do we find β?

Define cost function as the log (-negative) of $P(D|\beta)$

$$C(X,\beta) = -\sum_i \{y_i \log[P(y_i=1)] + (1-y_i)\log[1-P(y_i=1)]\}$$

$$P(t_i) = P(\beta_0 + \beta_1 x_i) = \frac{e^{t_i}}{1+e^{t_i}}$$

$$\frac{\partial C(X\beta)}{\partial \beta_j} = 0$$

$$\frac{\partial C}{\partial \beta} = -X^T(y - p) \in \mathbb{R}^P$$

$$p = [p_0, p_1 \ldots p_{n-1}]$$

$y, p \in \mathbb{R}^n \quad X \in \mathbb{R}^{n \times p}$

Define W with diagonal matrix elements only $p_i(1-p_i)$

$$\frac{\partial^2 C}{\partial \beta \partial \beta^T} = \underbrace{X^T W X}_{\text{always positive definite}} \in \mathbb{R}^{p \times p}$$

Need to solve

$$\boxed{X^T(y - p)} = 0 \qquad \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

non-linear in the unknown parameters β

OLS $\quad \beta = (X^T X)^{-1} X^T y$

Family of gradient descent methods.

$f(\beta) = 0$ Newton's method.

$$\frac{\partial C(\beta)}{}$$

$\partial\beta$

iterations k

$$\beta_{k+1} = \beta_k - f(\beta_k)/f'(\beta_k)$$

$$\|\beta_{k+1} - \beta_k\|_2 < \varepsilon \approx 10^{-10}$$

$$\beta_{k+1} = \beta_k - \underbrace{\left(\frac{\partial^2 C}{\partial\beta\,\partial\beta^T}\right)^{-1}_{\beta=\beta_k}} \times \underline{\left(\frac{\partial C}{\partial\beta}\right)_{\beta=\beta_k}}$$

$$\beta_{k+1} = \beta_k - \underline{\gamma_k}\left(\frac{\partial C}{\partial\beta}\right)_{\beta=\beta_k}$$

Learning rate