Lecture 3

DATA 558

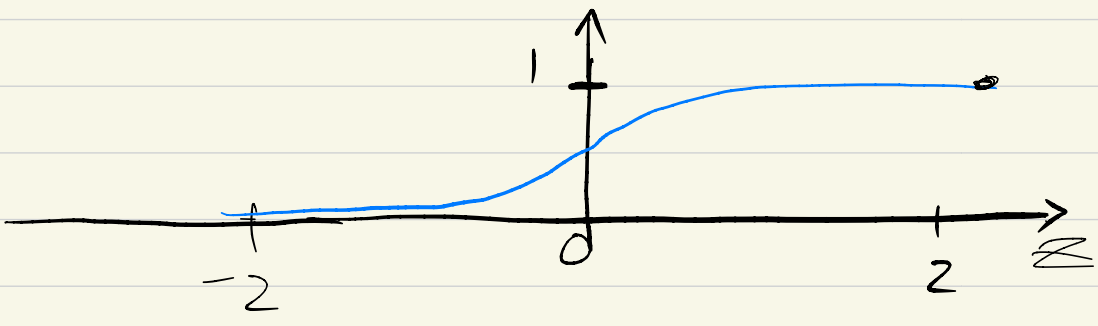
$$\min_{\beta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + e^{-y_i x_i^T \beta}\right) + \lambda \|\beta\|_2$$

$$\mathbb{P}(Y=1 \mid x; \beta) = g(x^T \beta)$$

$$g(z) = 1 / \left(1 + e^{-z}\right)$$

g is called logistic function sigmoid

STAT ML



2020

$$\mathbb{P}(Y=+1 \mid x;\beta) + \mathbb{P}(Y=-1 \mid x;\beta) = 1$$

exclusive class membership

$$\mathbb{P}(Y=-1 \mid x;\beta) = 1 - \mathbb{P}(Y=+1 \mid x;\beta)$$

$$\log\left(\frac{\mathbb{P}(Y=+1 \mid x;\beta)}{\mathbb{P}(Y=-1 \mid x;\beta)}\right) = x^T\beta$$

$g(z)$ nice property

$$1 - g(z) = g(-z)$$

$$g(-z) = \frac{1}{1 + e^{-(-z)}}$$

$$= \frac{1}{1 + e^{+z}}$$

$$= \frac{e^{-z} \cdot 1}{e^{-z}(1 + e^{+z})}$$

$$= \frac{e^{-z}}{e^{-z} + \underbrace{e^{-z+z}}_{1}}$$

$$= \frac{e^{-z}}{e^{-z} + 1}$$

$$1 - g(+z) = g(-z)$$

$$g(-z) = \frac{e^{-z}}{1 + e^{-z}}$$

$$= \frac{e^{-z} + 1 - 1}{1 + e^{-z}}$$

$$= \frac{e^{-z} + 1}{1 + e^{-z}} - \frac{1}{1 + e^{-z}}$$

$$= 1 - \frac{1}{1 + e^{-z}}$$

$$g(-z) = 1 - g(+z)$$

$$\mathbb{P}(Y=+1 \mid x;\beta)$$

$$= g(z) = g(x^T\beta)$$

$$\mathbb{P}(Y=-1 \mid x;\beta)$$

$$= 1 - g(z)$$

$$= g(-z) = g(-x^T\beta)$$

Un-regularized Maximum Likelihood Estimator for Logistic Regression

Likelihood $\mathcal{L}$

$$\mathcal{L}\left((x_1, y_1), \dots, (x_n, y_n)\right)$$

$$= \prod_{i=1}^{n} p(y_i \mid x_i ; \beta)$$

$$\uparrow$$

$$p(y_1 \mid x_1 ; \beta) \times \cdots \times p(y_n \mid x_n ; \beta)$$

$$\max_{\beta \in \mathbb{R}^d} \mathcal{L}(\beta)$$

$$\min_{\beta \in \mathbb{R}^d} -\log\left(\mathcal{L}(\beta)\right).$$

$$-\log\left(\prod_{i=1}^{n} p(y_i \mid x_i, \beta)\right)$$

$$= -\sum_{i=1}^{n} \log\left(p(y_i \mid x_i, \beta)\right)$$

$$\log(a \times b) = \log(a) + \log(b)$$

$$-\log\left(\prod_{i=1}^{n} \dots\right)$$

$$= \sum_{i=1}^{n} -\log p(y_i \mid x_i, \beta)$$

$$= \sum_{i=1}^{n} -\log g(y_i \cdot x_i^T \beta)$$

$$p(y=+1 \mid x_i, \beta) = g(x_i^T \beta) = g(+1 \cdot x_i^T \beta)$$

$$p(y=-1 \mid x_i, \beta) = 1 - g(x_i^T \beta)$$

$$= g(-x_i^T \beta) = g(-1 \cdot x_i^T \beta)$$

$$\mathbb{P}(Y=y \mid x;\beta) = g(y \cdot x^T\beta)$$

where $y=+1$ or $y=-1$

$$-\log(\text{Likelihood})$$

$$= \sum_{i=1}^{n} -\log\left(g(y_i \cdot x_i^T\beta)\right) \quad \left(\text{all terms write that way}\right)$$

$$= \sum_{i=1}^{n} (-)\log\left(\frac{1}{1+e^{-y_i \cdot x_i^T\beta}}\right)$$

but $\log\left(\frac{1}{a}\right) = -\log(a)$

$$= \sum_{i=1}^{n} \underbrace{(-1)\times(-1)}_{+1} \cdot \log\left(1+e^{-y_i x_i^T\beta}\right)$$

$$\frac{1}{n}\sum_{i=1}^{n} \log\left(1+e^{-y_i \cdot x_i^T\beta}\right)$$

Empirical risk / loss

$$\min_{\beta \in \mathbb{R}^d} f(\beta) \quad \left[\frac{1}{n} \sum_{i=1}^{n} \log\left(1 + e^{-y_i x_i^T \beta}\right) + \lambda \|\beta\|_2^2 \right]$$

Gradient descent algorithm

$$\beta_{t+1} = \beta_t - \eta_t \nabla_{\beta_t} F$$

η: eta

Single term $\frac{\partial}{\partial \beta} \log\left(1 + e^{-y \cdot x^T \beta}\right)$

$$f(\cdot) = \log(\cdot)$$

$$g(\cdot) = 1 + \exp(-\cdot)$$

$$h(\cdot) = y \cdot x^T \beta$$

$$f(g(h(\beta)))$$

$$f((g \circ h)(\beta))$$

$$(f \circ g \circ h)' = (g \circ h)' \cdot f'((g \circ h))$$

$$(g \circ h)' = h' \cdot g'(h)$$

$$(f \circ g \circ h)' = h' \cdot g'(h) \cdot f'(g \circ h)$$

$$\left(1 + \exp(-bla)\right)' = -\exp(-bla)$$

$$\frac{\partial}{\partial bla}\left(v^T bla\right) = v$$

$$h' = y \cdot x$$

$$g'(h) = -1 \cdot e^{-(yx)^T\beta}$$

$$f'(g \circ h) = \frac{1}{1 + e^{-y \cdot x^T\beta}}$$

$$\frac{\partial}{\partial\beta} \log\left(1 + e^{-y \cdot x^T\beta}\right)$$

$$= -yx \cdot \frac{e^{-y \cdot x^T\beta}}{1 + e^{-y \cdot x^T\beta}}$$

$$\frac{\partial}{\partial \beta} \left\{ \log\left(1 + e^{-y_i \cdot x_i^T \beta}\right) \right\}$$

$$= -y_i \cdot x_i \cdot \frac{e^{-y_i x_i^T \beta}}{1 + e^{-y_i x_i^T \beta}}$$

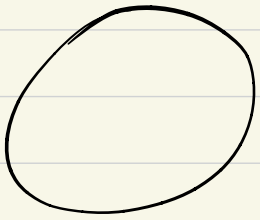
$$1 - p_i = \frac{e^{-y_i x_i^T \beta}}{1 + e^{-y_i x_i^T \beta}}$$

$$\frac{\partial}{\partial \beta} \{ \dots \} = -y_i x_i \cdot (1 - p_i)$$

log-sum-exp

$$\frac{1}{n}\sum_{i=1}^{n} \log\left(1 + e^{-y_i x_i^T \beta}\right)$$

$$= \frac{1}{n} \; ? \; \underline{Y} \quad \underline{X} \; ?$$



$$\frac{1}{n}\sum_{i=1}^{n} \left(y_i - x_i^T \beta\right)^2$$

$$= \frac{1}{n} \left\| \underline{Y} - \underline{X}\beta \right\|_2^2$$

$$Q = Id - Diag\left([p_1, \dots, p_n]\right)$$

p_i as defined earlier

$$Q = \begin{pmatrix} 1-p_1 & & \\ & \ddots & \\ & & 1-p_n \end{pmatrix}$$

$$\min_{\beta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + e^{-y_i x_i^T \beta}\right) R(\beta)$$

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

$$\frac{\partial}{\partial \beta} R(\beta) = \frac{1}{n} \sum_{i=1}^{n} -y_i x_i (1-p_i)$$

$$= -\frac{1}{n} Q Y X$$

$$\frac{\partial}{\partial \beta} \left\{ \lambda \|\beta\|_2^2 \right\} = ?$$

$$\frac{\partial f}{\partial \beta} = \boxed{-\frac{1}{n} Q Y X + 2\lambda \beta}$$

Regularized logistic regression

Penalized Maximum Likehood predictor

Mathematical derivation

Gradient Computation

Smoothness constant L

$$\|\nabla F(\beta) - \nabla F(\beta')\|_2^2 \leq L \|\beta - \beta'\|_2$$

For all $\beta, \beta' \in \mathbb{R}^d$

Gross estimate of L

Pick two β, β' at random.

$$L_{est} \simeq \frac{\|\nabla F(\beta) - \nabla F(\beta')\|_2^2}{\|\beta - \beta'\|_2}$$

quick

Better estimate of L

$L_{est} \simeq$ maximum eigenvalue of

$$\left\{ \left(\frac{1}{m} X_{sub}^T X_{sub} \right) + \lambda \, Id \right\}$$

Python largest-eigh.

slow

Adaptive step-size rule

Back-tracking line-search

Init $\eta_t = \eta_{t-1}$ if $t \geq 1$

$$\eta_0 = \frac{1}{L_{est}}$$

Iterate $\eta_t = \gamma \eta_t$

until condition is satisfied

$$F(\beta_{t+1}) = F(\beta_t - \eta_t \nabla f(\beta_t))$$

$$\leq F(\beta_t) - \alpha \eta_t \|\nabla f(\beta_t)\|_2^2$$

where $\alpha = 0.5$ and $\gamma = 0.8$

Fast Gradient Method

Init $\beta_0 = 0$; $\theta_0 = 0$

Iterate Find η_t by backtracking line-search

$$\beta_{t+1} = \theta_t - \eta_t \nabla_\theta F(\theta_t)$$

MOMENTUM $$\theta_{t+1} = \beta_{t+1} + \frac{t}{t+3}(\beta_{t+1} - \beta_t)$$

$(\beta_t)_{t \geq 1}$

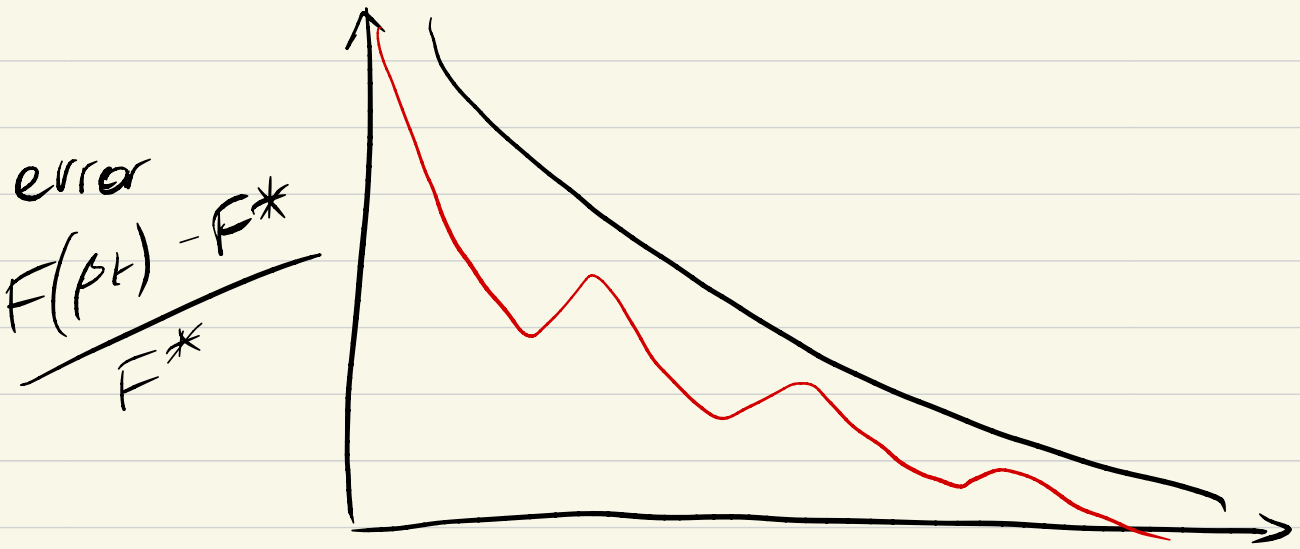
$(\theta_t)_{t \geq 1}$

two sequences of iterates of the same time

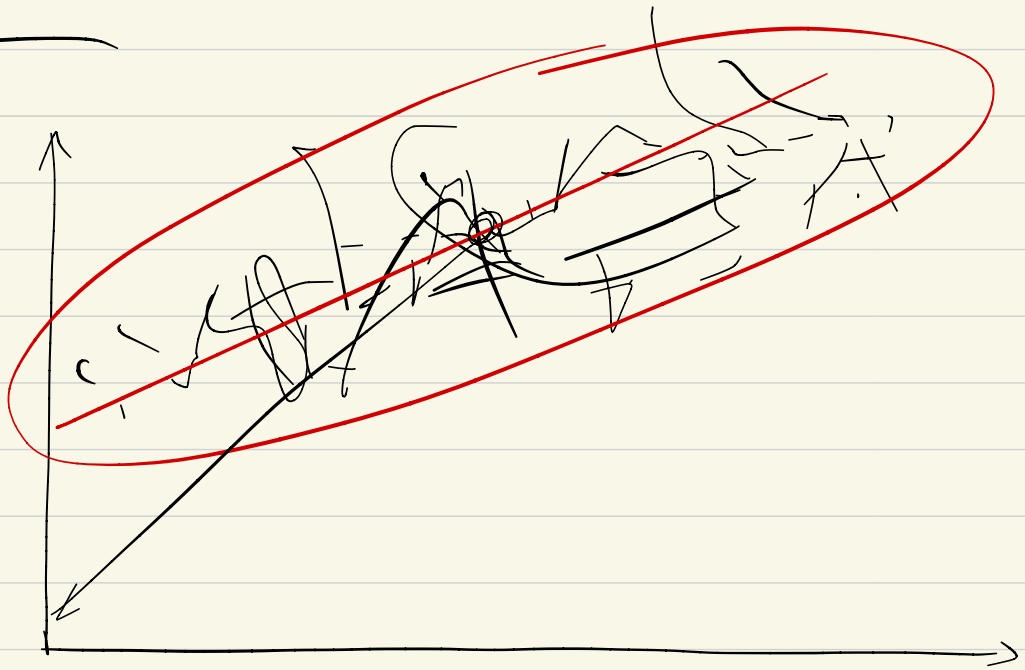
$$O\left(\frac{1}{t}\right) \longrightarrow O\left(\frac{1}{t^2}\right)$$

Nesterov accelerated gradient method

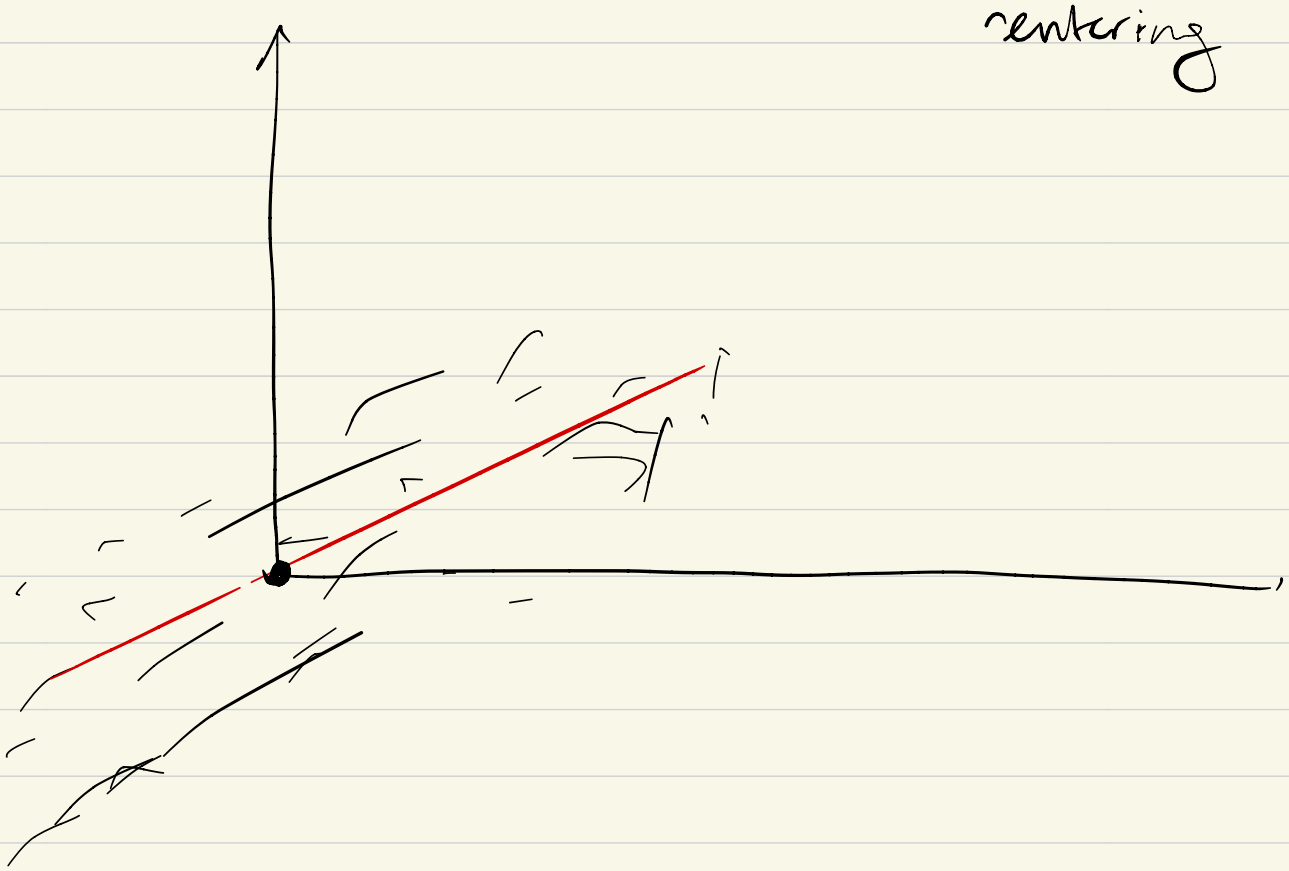
Polyak momentum



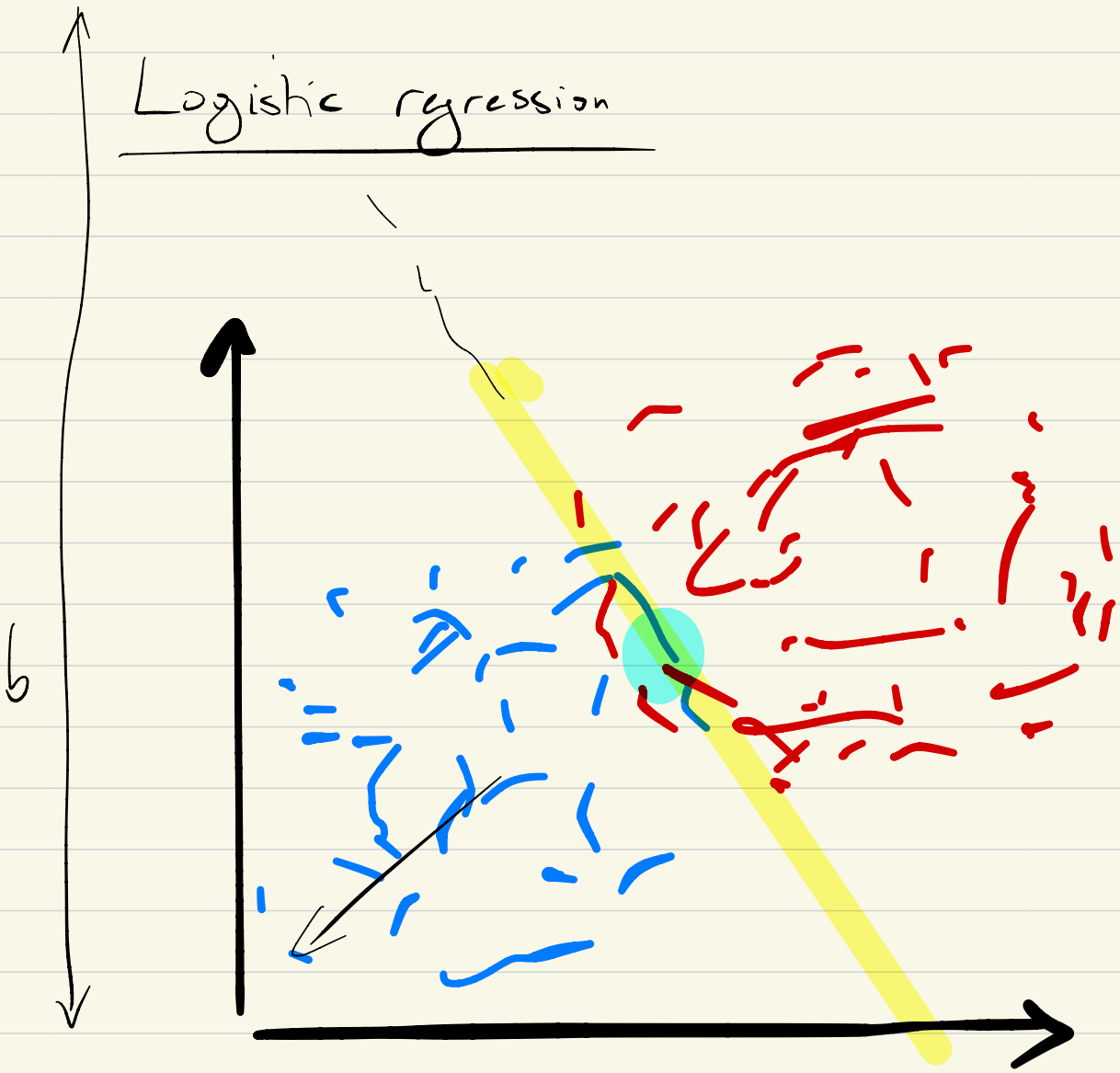
Regression



centering



Logistic regression

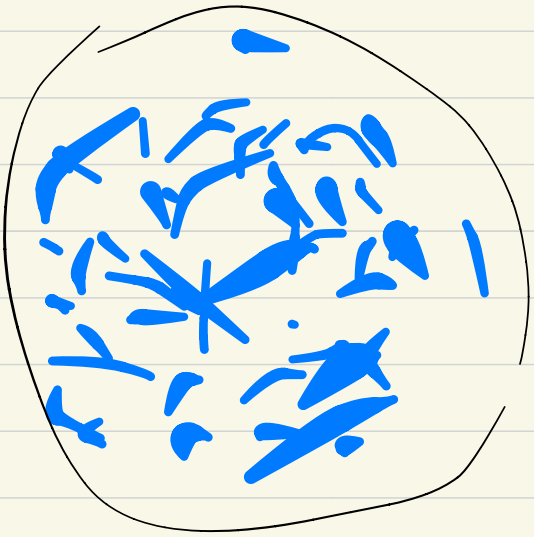


Training data

spam

non-spam

Testing data



$$\tilde{x} = \begin{bmatrix} & \\ & \\ & \end{bmatrix} \quad \begin{bmatrix} 1 \end{bmatrix}$$

$$\tilde{\beta} = \begin{bmatrix} & \\ & \\ & \end{bmatrix} \quad \begin{bmatrix} \uparrow \end{bmatrix}$$

intercept

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

actual features

$$+ \; 1 \times \text{intercept}$$

$$= x^T \beta + \text{intercept}$$

$$\|\tilde{\beta}\|_2^2 = \sum_{j=1}^{d} \tilde{\beta}_j^2 + \text{intercept}^2$$