- Thanks for staying/auditing

Ridge Regression
MAP
Kernels

- Ass 2 due Monday
  (Tutorial tomorrow)
- Graded Ass1 due Monday

- Notes on webpage:
  - Gaussian MLE
  - Solving least squares
- You do not need to know linear programming
- but you do need to know the trick

Monday

due Monday

## Converting to constrained problem

- non-smooth optimization is hard
- smooth constrained optimization often simpler

Trick to convert (1 of many):

$\max\{x_1, x_2\} = |x|$

If we have $\min_x \max_i \{f_i(x)\}$

equivalent

$\min_{x,v} v$ subject to $v \geq \max_i \{f_i(x)\}$

$\Leftrightarrow$

$v \geq f_i(x), \forall i$

Ch

Orig

x=

# Change of Basis

Original data:

$$X = \begin{bmatrix} 1 \\ .5 \\ 2 \\ \vdots \\ 10 \end{bmatrix}$$

make new X

"$\Phi(X)$"

$\bar{x}_i \rightsquigarrow \phi(\bar{x}_j)$

template

$$\begin{bmatrix} 1 & x & x^2 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & .5 & .25 \\ 1 & 2 & 4 \\ \vdots & \vdots & \vdots \\ 1 & 10 & 100 \end{bmatrix}$$

new design matrix

(then just business as usual)

$= 1 \times 1$

late

$x^2$ ]

1 ⎫

.25 ⎪

4 ⎪

⋮ ⎪

100 ⎭

design

trix

# Cross-Validation

- large 'k': less chance of overestimating test error, but higher variance + running time.

- idea of repeating CV with different random partitions: lower variance.

# Ridge Regression

MAP

kernels

## Ridge Regression

$$f(\bar{w}) = \frac{1}{2}\|X\bar{w} - Y\|^2 + \frac{\lambda}{2}\|\bar{w}\|^2$$

$\nabla f(\bar{w}) = X^T$

What? Shink $\bar{w}_i$ toward 0

Why?

- solution is unique
- avoids overfitting due to huge values
- improves conditioning

"Magic": do a huge basis expansion

$D^2 f$

ession

$\frac{\lambda}{2}||\bar{w}||^2$

$$\nabla f(\bar{w}) = X^T X \bar{w} - X^T Y + \lambda \bar{w}$$

$$= 0 \quad \text{(at stationary point)}$$

$$(X^T X + \lambda I)\bar{w} = X^T Y$$

$$\therefore \bar{w} = (X^T X + \lambda I)^{-1} X^T Y$$

$\bar{w} = X^T(XX^T$

$$\nabla^2 f(w) = X^T X + \lambda I \succ 0$$

"positive definite"

solution is unique

$\frac{1}{N}X^T$

$XX$

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

positive
definite"

$X^TX$ "scatter matrix"

$\frac{1}{N}X^TX$ "covariance MLE"

$XX^T$ "Gram matrix"

# MAP estimation

MLE: $\underset{h \in H}{\text{argmax}}$

MAP: $\underset{h \in H}{\text{argmax}}$

$p(D|$

(poste

$p(h|$

ation

max
H.

max
H

$f(x) \propto x$

$= cx$

$f(x) \propto \frac{y^2+\alpha^3}{\gamma} x$

"Is model h reasonable"

(likelihood) (prior)

$p(D|h)$

(posterior)

$p(h|D) \propto p(D|h)p(h)$

$\sum_x p(x) = 1$

"exists constant, not dependent h,
s.t. $\sum_h p(h|D) = 1$"

Ridge Regression

MAP

kernels

## Ridge Regression

$$\underbrace{\frac{1}{2}\|X\bar{w}-Y\|^2}_{\text{log-likelihood}} + \underbrace{\frac{\lambda}{2}\|\bar{w}\|^2}_{\text{prior}}$$

$$-\log p(h|D) = -\log p(D|h) - \log p(h) + \text{const}$$

$$= \sum_{i=1}^{N} \frac{1}{2}(y_i - \bar{x}_i^T w)^2 + \frac{1}{2\sigma^2}\bar{w}^T\bar{w}$$

"precision" $\lambda = \frac{1}{\sigma^2}$

$$\frac{\lambda}{2}\|\bar{w}\|^2$$

$y_i | x_i$

$w_i | \theta$

$p(w_i) =$

$$y_i \mid \bar{x}_i, \bar{w} \sim N(\bar{w}^T \bar{x}_i, 1)$$

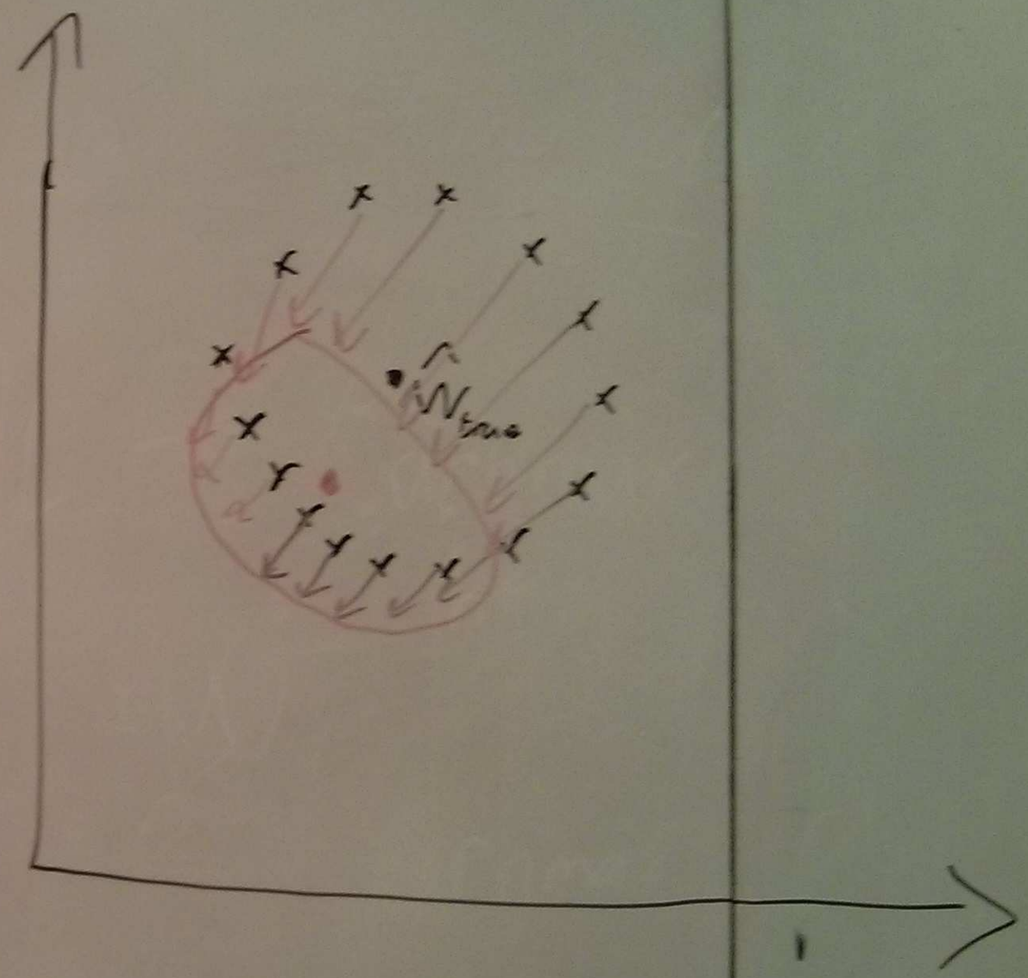
$$w_i \mid \theta^2 \sim N(0, \theta^2)$$

$$p(w_i) = \frac{1}{\theta\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(w_i - 0)^2}{\theta^2}\right)$$

$$\log p(w_i) = -\frac{1}{2\theta^2} w_i^2 \underbrace{- \log(\theta) - \log(\sqrt{2\pi})}_{\text{const.}} + \text{const.}$$

$$\log(ab) = \log$$

$$\log\left(\frac{a}{b}\right) = \log$$



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

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

(124)

Ridge Regression
MAP
Kernels

## Kernels

We assume $\bar{x}_i \in \mathbb{R}^d$

What if don't know representation?

E.g. "the cat is back"

"the black cat is back"

LCS

Edit distance

"Kernel

$k(\bar{x}_i$

$\bar{x}_i \in$

"the c

Typica

"Kernel function"

$k(\bar{x}_i, \bar{x}_j) \longmapsto \mathbb{R}$, often a similarity.

$\bar{x}_i \in \mathcal{X}$

"the cat is back" $\in \mathcal{X}$

Typically, $k(\bar{x}_i, \bar{x}_j) \geq 0$

$k(\bar{x}_i, \bar{x}_j) = k(\bar{x}_j, \bar{x}_i)$

Example

vecto

"Linear"

"Poly"

"RBF"

How to use?

$$\phi(\bar{x}_i) = \left[ k(\bar{x}_i, \bar{x}_1) \;\; k(\bar{x}_i, \bar{x}_2) \;\; \cdots \;\; k(\bar{x}_i, \bar{x}_N) \right]$$

$$\left[ k(\bar{x}_i, \bar{z}_1) \;\; k(\bar{x}_i, z_2) \;\; \cdots \;\; k(\bar{x}_i, z_M) \right]$$

$$\underline{\Phi}(X) = \begin{bmatrix} k(\bar{x}_1, \bar{x}_1) & k(\bar{x}_1, \bar{x}_2) & \cdots \\ k(\bar{x}_2, \bar{x}_1) & \ddots & \\ \vdots & & \ddots \\ & & & k(\bar{x}_N, \bar{x}_N) \end{bmatrix} \rightarrow \text{Gram matrix}$$

linear: $\underline{\Phi}(X) = X X^T$

## Examples

vectors:

"Linear": $\bar{x}_i^T \bar{x}_j$

"Poly": $(\bar{x}_i^T \bar{x}_j + \alpha)^m$

"RBF": $\exp\left(-\frac{\|\bar{x}_i - \bar{x}_j\|^2}{2\sigma^2}\right)$

Section 14.2:

- string kernels
- pyramid match kernels

Ridge Reg.

MAP

Kernels

Kernel Trick

Some kernels have explicit feature map (inner product)

$$k(x,z) = (x^T z)^2$$
$$= (x_1 z_1 + x_2 z_2)^2$$
$$= x_1^2 z_1^2 + 2 x_1 x_2 z_1 z_2 + x_2^2 z_2^2$$
$$= (x_1^2, \sqrt{2}\, x_1 x_2, x_2^2)^T (z_1^2, \sqrt{2}\, z_1 z_2, z_2^2)$$
$$= \phi(x)^T \phi(z)$$

$w = w_{MAP}$

Test time

$x = x$

$= x$

$= k$

$$\bar{w}_{MAP} = X^T (XX^T + \lambda I)^{-1} y$$

Test time

$$\hat{y} = \hat{X}\, \bar{w}_{MAP}$$

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

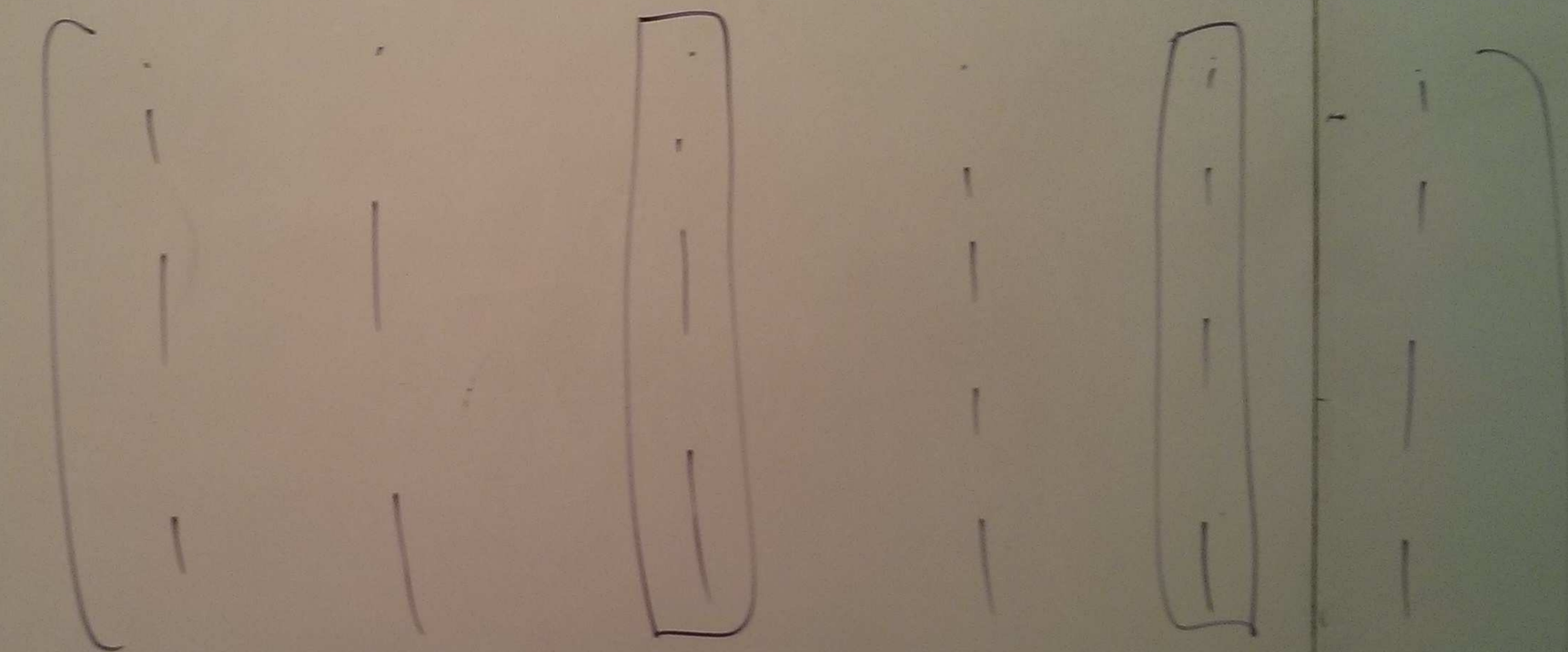
$k(\hat{X}, X)$ Gram matrix $k(X, X)$

$$= k(\hat{X}, X)\left(k(X, X) + \lambda I\right)^{-1} y$$

"Gaussian Process"

# Feature Selection

What if we don't know which $x_i^j$ are relevant?

$X =$ 

Search over features:

- NP-hard
- fine for 16 vars, 1 million vars