

Regularized Linear Regression, Nonlinear Regression

Aarti Singh

Machine Learning 10-315

Oct 30, 2019



MACHINE LEARNING DEPARTMENT



Least Squares and M(C)LE

Intuition: Signal plus (zero-mean) Noise model

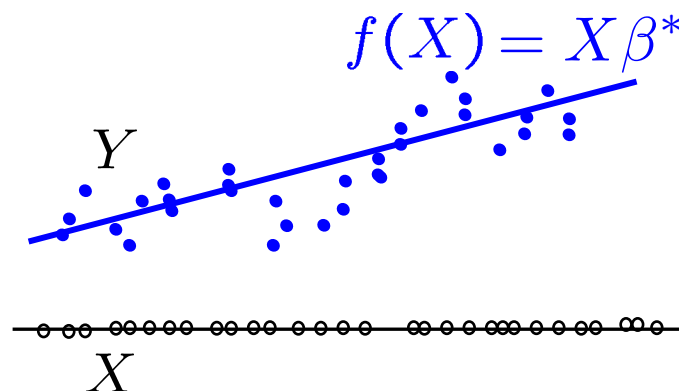
$$Y = f^*(X) + \epsilon = X\beta^* + \epsilon$$

$$\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I}) \quad Y \sim \mathcal{N}(X\beta^*, \sigma^2 \mathbf{I})$$

$$\hat{\beta}_{\text{MLE}} = \arg \max_{\beta} \underbrace{\log p(\{Y_i\}_{i=1}^n | \beta, \sigma^2, \{X_i\}_{i=1}^n)}_{\text{Conditional log likelihood}}$$

Conditional log likelihood

$$= \arg \min_{\beta} \sum_{i=1}^n (X_i \beta - Y_i)^2 = \hat{\beta}$$



Least Square Estimate is same as Maximum Conditional Likelihood Estimate under a Gaussian model !

Regularized Least Squares and M(C)AP

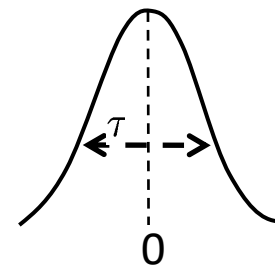
What if $(\mathbf{A}^T \mathbf{A})$ is not invertible ?

$$\hat{\beta}_{\text{MAP}} = \arg \max_{\beta} \underbrace{\log p(\{Y_i\}_{i=1}^n | \beta, \sigma^2, \{X_i\}_{i=1}^n)}_{\text{Conditional log likelihood}} + \underbrace{\log p(\beta)}_{\text{log prior}}$$

I) Gaussian Prior

$$\beta \sim \mathcal{N}(0, \tau^2 \mathbf{I})$$

$$p(\beta) \propto e^{-\beta^T \beta / 2\tau^2}$$



$$\hat{\beta}_{\text{MAP}} = \arg \min_{\beta} \sum_{i=1}^n (Y_i - X_i \beta)^2 + \lambda \|\beta\|_2^2$$

constant(σ^2, τ^2)

Ridge Regression

$$\hat{\beta}_{\text{MAP}} = (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T \mathbf{Y}$$

Regularized Least Squares and M(C)AP

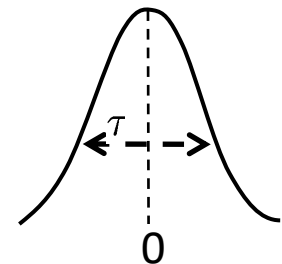
What if $(\mathbf{A}^T \mathbf{A})$ is not invertible ?

$$\hat{\beta}_{\text{MAP}} = \arg \max_{\beta} \underbrace{\log p(\{Y_i\}_{i=1}^n | \beta, \sigma^2, \{X_i\}_{i=1}^n)}_{\text{Conditional log likelihood}} + \underbrace{\log p(\beta)}_{\text{log prior}}$$

I) Gaussian Prior

$$\beta \sim \mathcal{N}(0, \tau^2 \mathbf{I})$$

$$p(\beta) \propto e^{-\beta^T \beta / 2\tau^2}$$



$$\hat{\beta}_{\text{MAP}} = \arg \min_{\beta} \sum_{i=1}^n (Y_i - X_i \beta)^2 + \lambda \|\beta\|_2^2$$

\downarrow
constant(σ^2, τ^2)

Ridge Regression

Prior belief that β is Gaussian with zero-mean biases solution to “small” β

Regularized Least Squares and M(C)AP

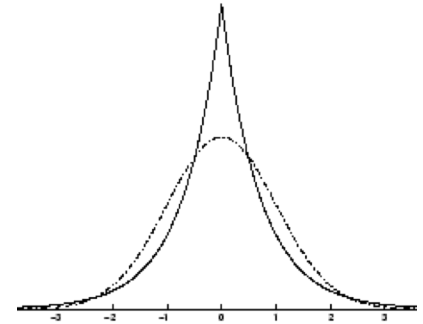
What if $(\mathbf{A}^T \mathbf{A})$ is not invertible ?

$$\hat{\beta}_{\text{MAP}} = \arg \max_{\beta} \underbrace{\log p(\{Y_i\}_{i=1}^n | \beta, \sigma^2, \{X_i\}_{i=1}^n)}_{\text{Conditional log likelihood}} + \underbrace{\log p(\beta)}_{\text{log prior}}$$

II) Laplace Prior

$\beta_i \stackrel{iid}{\sim} \text{Laplace}(0, t)$

$$p(\beta_i) \propto e^{-|\beta_i|/t}$$



$$\hat{\beta}_{\text{MAP}} = \arg \min_{\beta} \sum_{i=1}^n (Y_i - X_i \beta)^2 + \lambda \|\beta\|_1$$

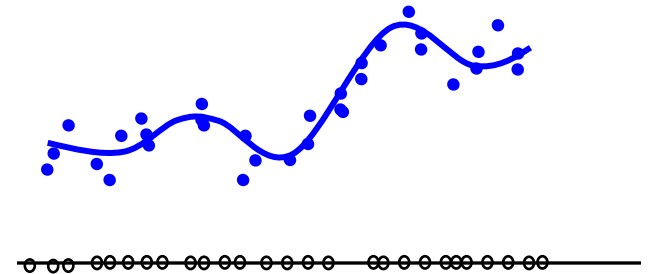
\downarrow
constant(σ^2, t)

Lasso

Prior belief that β is Laplace with zero-mean biases solution to “sparse” β

Beyond Linear Regression

Polynomial regression
Regression with nonlinear features



Kernelized Ridge Regression

Local Kernel Regression

Polynomial Regression

degree m

Univariate (1-dim) $f(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_m X^m = \mathbf{X}\beta$
case:

where $\mathbf{X} = [1 \ X \ X^2 \ \dots \ X^m]$, $\beta = [\beta_1 \ \dots \ \beta_m]^T$

$$\hat{\beta} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{Y} \text{ or } (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T \mathbf{Y} \quad \hat{f}_n(X) = \mathbf{X} \hat{\beta}$$

$$\text{where } \mathbf{A} = \begin{bmatrix} 1 & X_1 & X_1^2 & \dots & X_1^m \\ \vdots & & & \ddots & \vdots \\ 1 & X_n & X_n^2 & \dots & X_n^m \end{bmatrix}$$

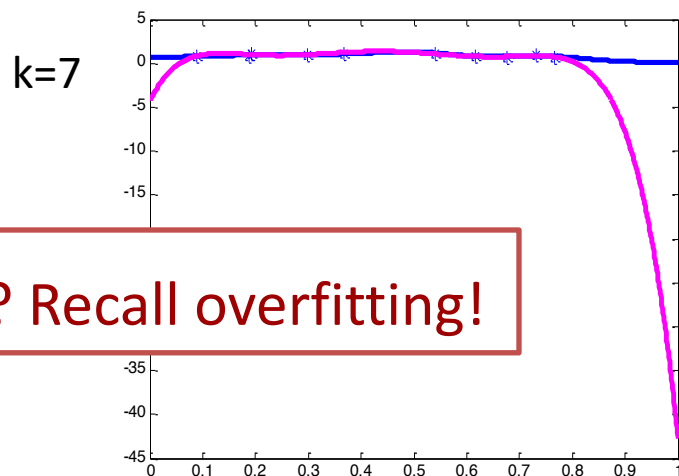
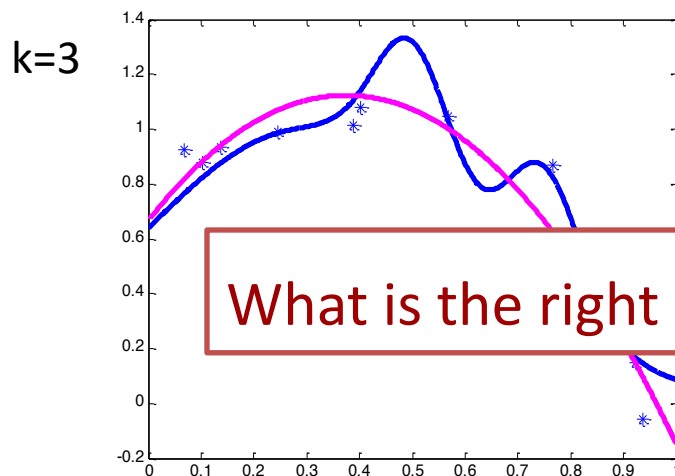
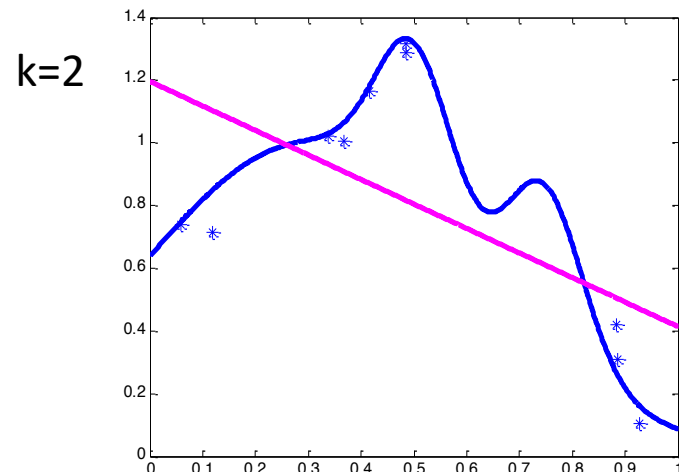
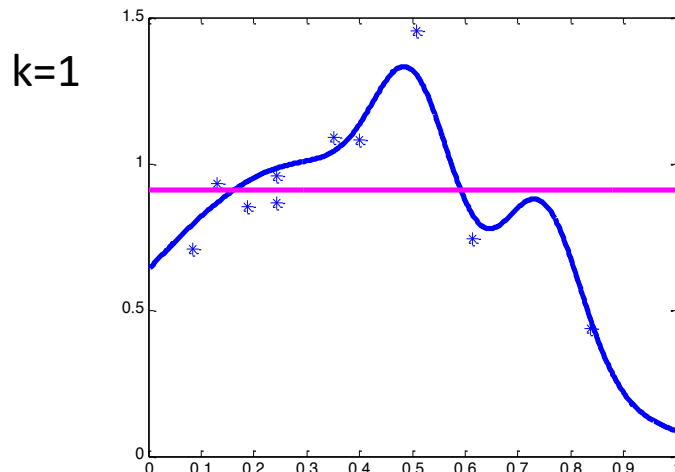
Multivariate (p-dim) $f(X) = \beta_0 + \beta_1 X^{(1)} + \beta_2 X^{(2)} + \dots + \beta_p X^{(p)}$
case:

$$+ \sum_{i=1}^p \sum_{j=1}^p \beta_{ij} X^{(i)} X^{(j)} + \sum_{i=1}^p \sum_{j=1}^p \sum_{k=1}^p X^{(i)} X^{(j)} X^{(k)}$$

+ ... terms up to degree m

Polynomial Regression

Polynomial of order k , equivalently of degree up to $k-1$

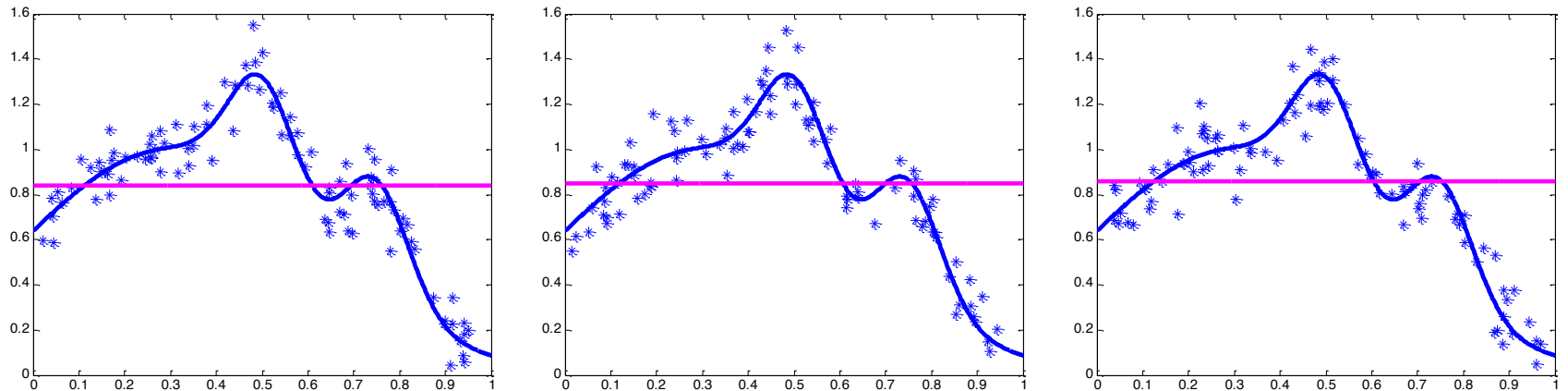


What is the right order? Recall overfitting!

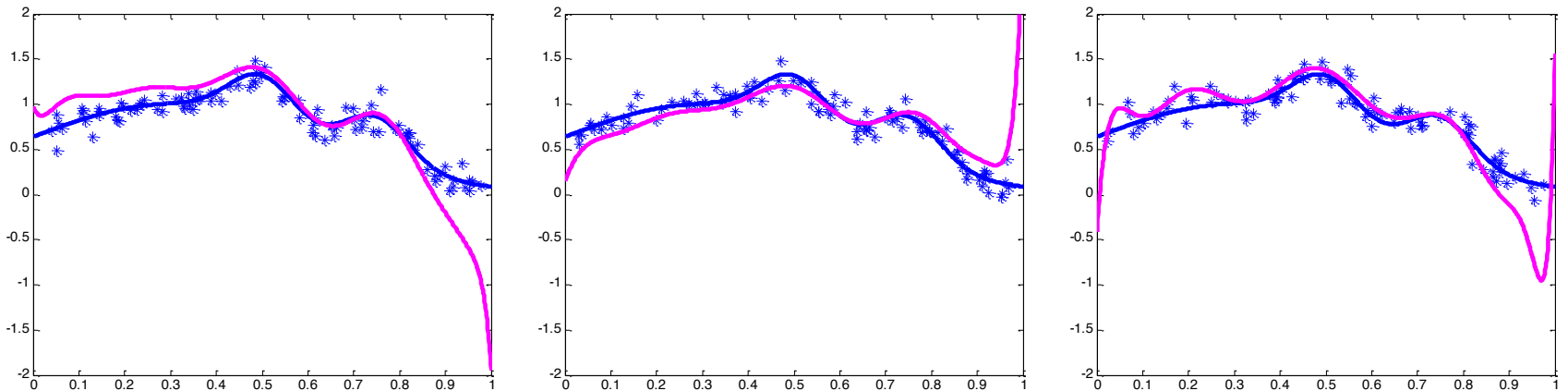
Bias – Variance Tradeoff

3 Independent training datasets

Large bias, Small variance – poor approximation but robust/stable



Small bias, Large variance – good approximation but unstable



Bias – Variance Decomposition

It can be shown that

$$E[(f(X) - f^*(X))^2] = \text{Bias}^2 + \text{Variance}$$

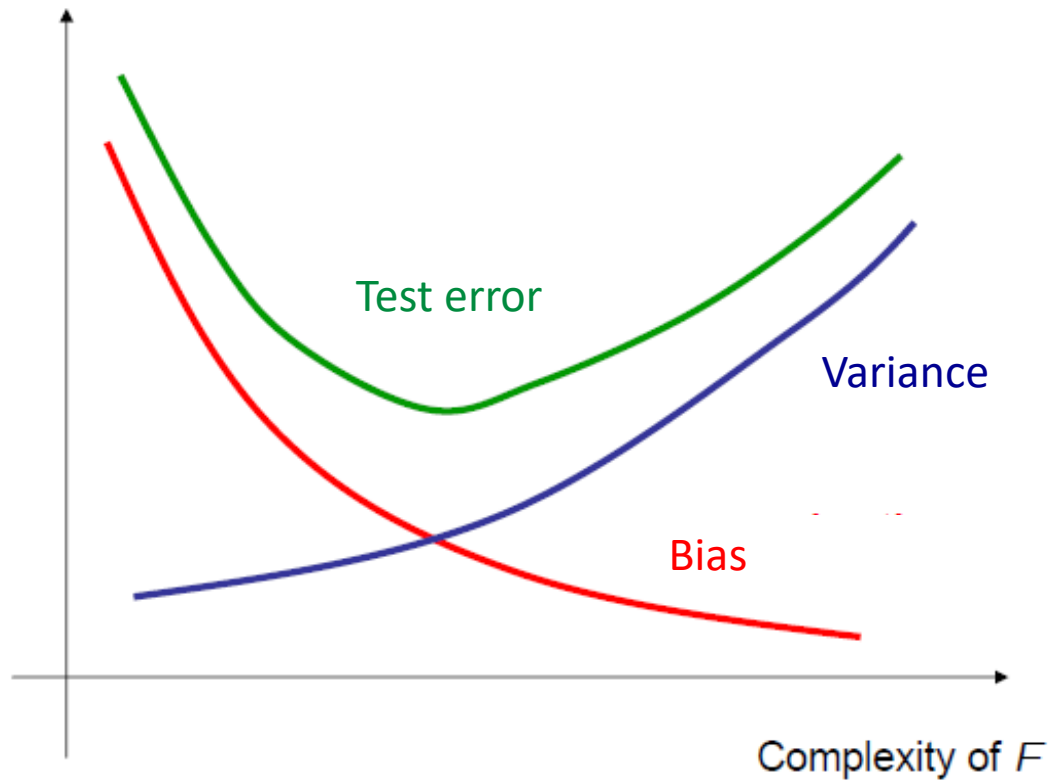
$$\text{Bias} = E[f(X)] - f^*(X)$$

How far is the model from
best model on average

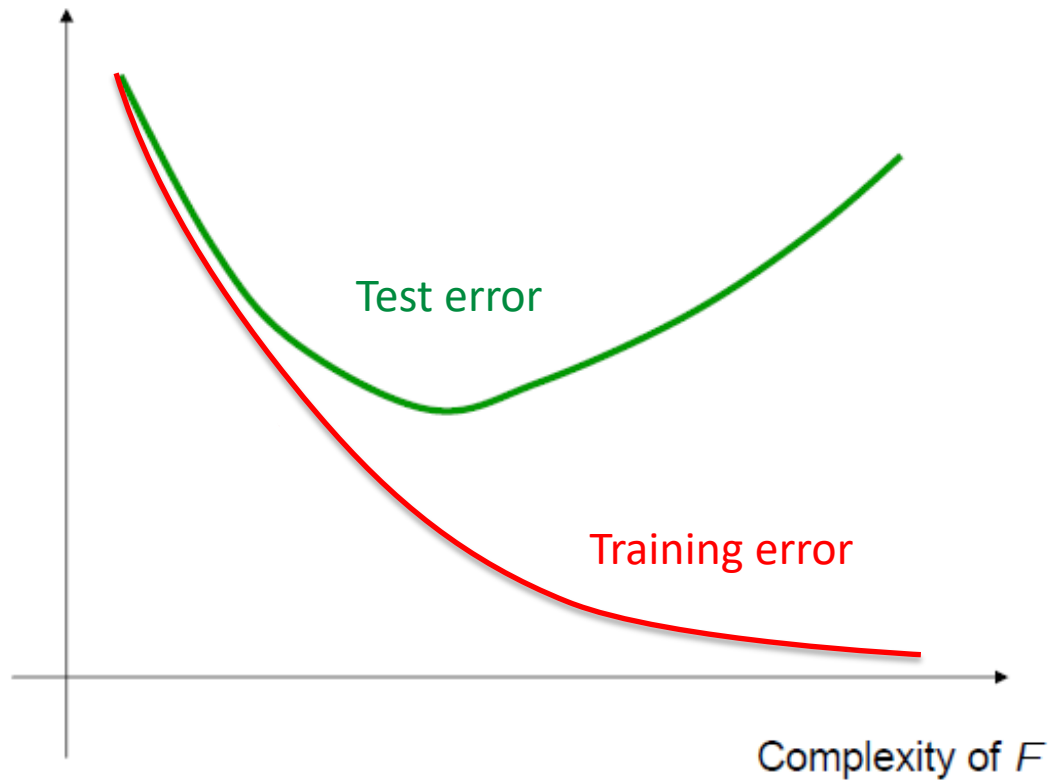
$$\text{Variance} = E[(f(X) - E[f(X)])^2]$$

How variable is the model

Effect of Model Complexity



Effect of Model Complexity

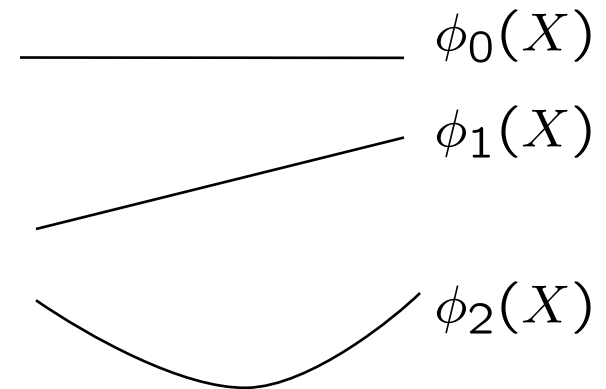


Regression with nonlinear features

$$f(X) = \sum_{j=0}^m \beta_j X^j = \sum_{j=0}^m \beta_j \phi_j(X)$$

Weight of
each feature

Nonlinear
features



In general, use any nonlinear features

e.g. e^X , $\log X$, $1/X$, $\sin(X)$, ...

$$\hat{\beta} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{Y}$$

or

$$(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T \mathbf{Y}$$

$$\mathbf{A} = \begin{bmatrix} \phi_0(X_1) & \phi_1(X_1) & \dots & \phi_m(X_1) \\ \vdots & & \ddots & \vdots \\ \phi_0(X_n) & \phi_1(X_n) & \dots & \phi_m(X_n) \end{bmatrix}$$

$$\hat{f}_n(X) = \mathbf{X} \hat{\beta}$$

$$\mathbf{X} = [\phi_0(X) \ \phi_1(X) \ \dots \ \phi_m(X)]$$

Can we use kernels?

Ridge regression (dual)

$$\min_{\beta} \sum_{i=1}^n (Y_i - X_i \beta)^2 + \lambda \|\beta\|_2^2 \quad \hat{\beta} = (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T \mathbf{Y}$$

Similarity with SVMs

Primal problem:

$$\begin{aligned} \min_{\beta, z_i} \quad & \sum_{i=1}^n z_i^2 + \lambda \|\beta\|_2^2 \\ \text{s.t.} \quad & z_i = Y_i - X_i \beta \end{aligned}$$

SVM Primal problem:

$$\begin{aligned} \min_{w, \xi_i} \quad & C \sum_{i=1}^n \xi_i + \frac{1}{2} \|w\|_2^2 \\ \text{s.t.} \quad & \xi_i = \max(1 - Y_i X_i w, 0) \end{aligned}$$

Lagrangian:

$$\sum_{i=1}^n z_i^2 + \lambda \|\beta\|_2^2 + \sum_{i=1}^n \alpha_i (z_i - Y_i + X_i \beta)$$

α_i – Lagrange parameter, one per training point

Ridge regression (dual)

$$\min_{\beta} \sum_{i=1}^n (Y_i - X_i \beta)^2 + \lambda \|\beta\|_2^2 \quad \hat{\beta} = (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T \mathbf{Y}$$

Dual problem:

$$\max_{\alpha} \min_{\beta, z_i} \sum_{i=1}^n z_i^2 + \lambda \|\beta\|^2 + \sum_{i=1}^n \alpha_i (z_i - Y_i + X_i \beta)$$

$\alpha = \{\alpha_i\}$ for $i = 1, \dots, n$

Taking derivatives of Lagrangian wrt β and z_i we get:

$$\beta = -\frac{1}{2\lambda} \mathbf{A}^T \alpha \quad z_i = -\frac{\alpha_i}{2}$$

Dual problem:
$$\max_{\alpha} -\frac{\alpha^T \alpha}{4} - \frac{1}{4\lambda} \alpha^T \mathbf{A} \mathbf{A}^T \alpha - \alpha^T \mathbf{Y}$$

n-dimensional optimization problem

Ridge regression (dual)

$$\min_{\beta} \sum_{i=1}^n (Y_i - X_i \beta)^2 + \lambda \|\beta\|_2^2 \quad \hat{\beta} = (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T \mathbf{Y}$$
$$= \mathbf{A}^T (\mathbf{A} \mathbf{A}^T + \lambda \mathbf{I})^{-1} \mathbf{Y}$$

Dual problem:

$$\max_{\alpha} -\frac{\alpha^T \alpha}{4} - \frac{1}{4\lambda} \alpha^T \mathbf{A} \mathbf{A}^T \alpha - \alpha^T \mathbf{Y} \quad \Rightarrow \hat{\alpha} = - \left(\frac{\mathbf{A} \mathbf{A}^T}{\lambda} + \mathbf{I} \right)^{-1} 2 \mathbf{Y}$$

can get back $\hat{\beta} = -\frac{1}{2\lambda} \mathbf{A}^T \hat{\alpha} = \mathbf{A}^T (\mathbf{A} \mathbf{A}^T + \lambda \mathbf{I})^{-1} \mathbf{Y}$

Weighted average of
training points

Weight of each training point (but typically not sparse)

Kernelized ridge regression

$$\hat{\beta} = (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T \mathbf{Y}$$

$$\hat{f}_n(X) = \mathbf{X} \hat{\beta}$$

Using dual, can re-write solution as:

$$\hat{\beta} = \mathbf{A}^T (\mathbf{A} \mathbf{A}^T + \lambda \mathbf{I})^{-1} \mathbf{Y}$$

How does this help?

- Only need to invert $n \times n$ matrix (instead of $p \times p$ or $m \times m$)
- More importantly, kernel trick!

$$\hat{f}_n(X) = \mathbf{K}_X (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{Y} \quad \text{where} \quad \begin{aligned} \mathbf{K}_X(i) &= \phi(X) \cdot \phi(X_i) \\ \mathbf{K}(i, j) &= \phi(X_i) \cdot \phi(X_j) \end{aligned}$$

Work with kernels, never need to write out the high-dim vectors

Kernelized ridge regression

$$\hat{f}_n(X) = \mathbf{K}_X(\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{Y} \quad \text{where} \quad \begin{aligned} \mathbf{K}_X(i) &= \phi(X) \cdot \phi(X_i) \\ \mathbf{K}(i, j) &= \phi(X_i) \cdot \phi(X_j) \end{aligned}$$

Work with kernels, never need to write out the high-dim vectors

Examples of kernels:

Polynomials of degree exactly d $K(\mathbf{u}, \mathbf{v}) = (\mathbf{u} \cdot \mathbf{v})^d$

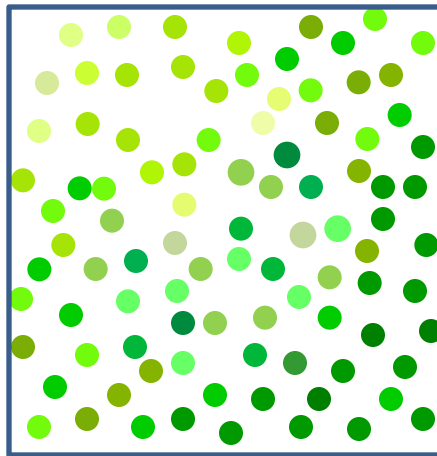
Polynomials of degree up to d $K(\mathbf{u}, \mathbf{v}) = (\mathbf{u} \cdot \mathbf{v} + 1)^d$

Gaussian/Radial kernels $K(\mathbf{u}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{u} - \mathbf{v}\|^2}{2\sigma^2}\right)$

Ridge Regression with (implicit) nonlinear features $\phi(X)$! $f(X) = \phi(X)\beta$

Local Kernel Regression

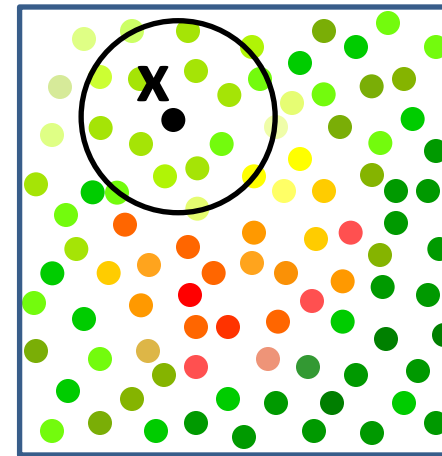
- What is the temperature in the room?



$$\hat{T} = \frac{1}{n} \sum_{i=1}^n Y_i$$

Average

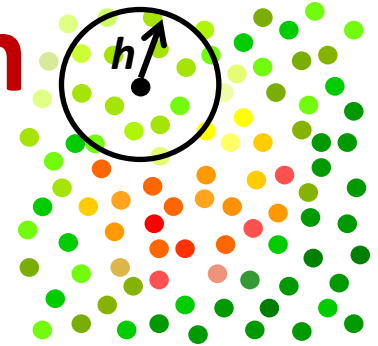
at location x ?



$$\hat{T}(x) = \frac{\sum_{i=1}^n Y_i \mathbf{1}_{\|X_i - x\| \leq h}}{\sum_{i=1}^n \mathbf{1}_{\|X_i - x\| \leq h}}$$

"Local" Average

Local Kernel Regression



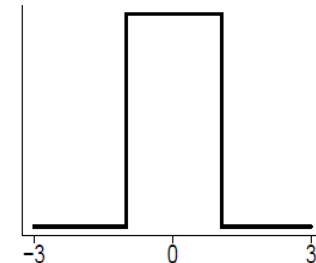
- Nonparametric estimator akin to kNN
- Nadaraya-Watson Kernel Estimator

$$\hat{f}_n(X) = \sum_{i=1}^n w_i Y_i \quad \text{Where} \quad w_i(X) = \frac{K\left(\frac{X - X_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{X - X_i}{h}\right)}$$

- Weight each training point based on distance to test point
- Boxcar kernel yields local average

boxcar kernel :

$$K(x) = \frac{1}{2}I(x),$$



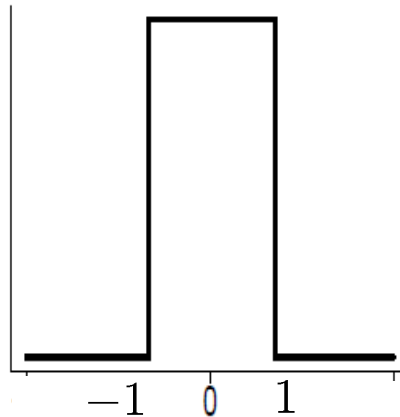
$$K(x) \geq 0,$$

$$\int K(x)dx = 1$$

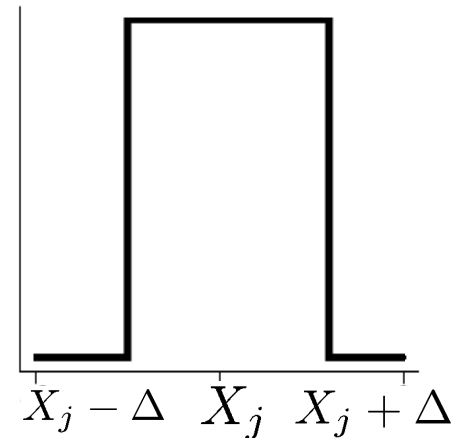
Kernels

boxcar kernel :

$$K(x) = \frac{1}{2}I(x),$$

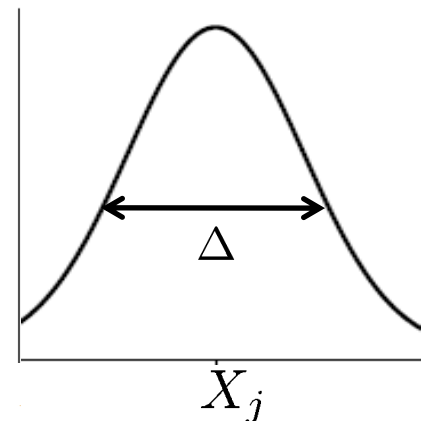
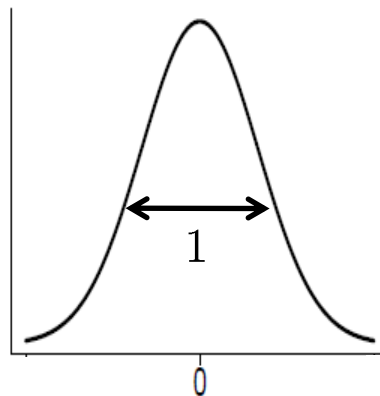


$$K\left(\frac{X_j - x}{\Delta}\right)$$



Gaussian kernel :

$$K(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$$



Choice of kernel bandwidth h

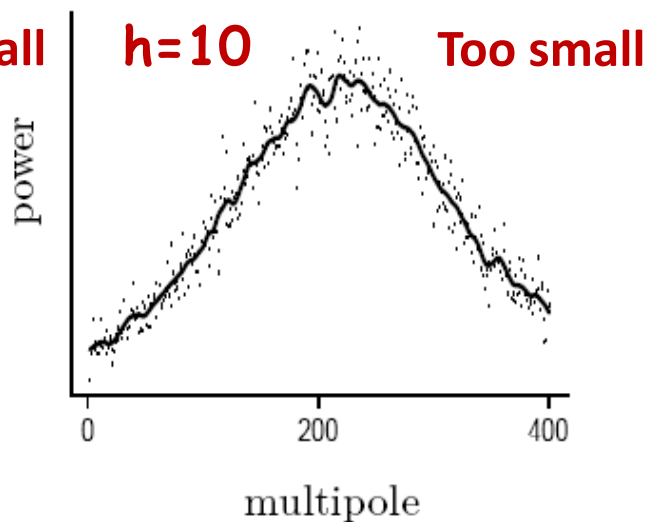
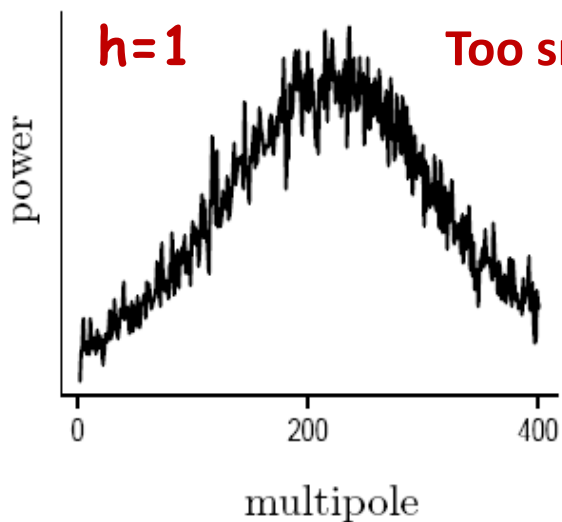
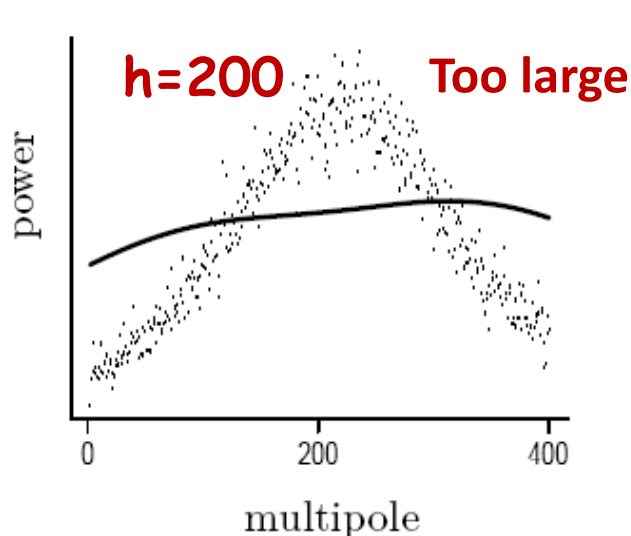
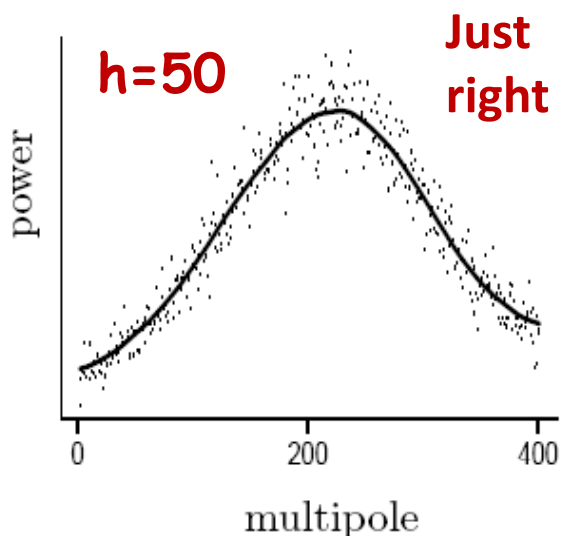


Image Source:
Larry's book – All
of Nonparametric
Statistics

Choice of kernel is
not that important



Kernel Regression as Weighted Least Squares

$$\min_f \sum_{i=1}^n w_i (f(X_i) - Y_i)^2 \qquad w_i(X) = \frac{K\left(\frac{X - X_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{X - X_i}{h}\right)}$$

Weighted Least Squares

Kernel regression corresponds to locally constant estimator obtained from (locally) weighted least squares

i.e. set $f(X_i) = \beta$ (a constant)

Kernel Regression as Weighted Least Squares

set $f(X_i) = \beta$ (a constant)

$$\min_{\beta} \sum_{i=1}^n w_i (\underbrace{\beta}_{\text{constant}} - Y_i)^2$$

$$w_i(X) = \frac{K\left(\frac{X - X_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{X - X_i}{h}\right)}$$

$$\frac{\partial J(\beta)}{\partial \beta} = 2 \sum_{i=1}^n w_i (\beta - Y_i) = 0$$

Notice that $\sum_{i=1}^n w_i = 1$

$$\Rightarrow \hat{f}_n(X) = \hat{\beta} = \sum_{i=1}^n w_i Y_i$$

Local Linear/Polynomial Regression

$$\min_f \sum_{i=1}^n w_i (f(X_i) - Y_i)^2$$

$$w_i(X) = \frac{K\left(\frac{X - X_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{X - X_i}{h}\right)}$$

Weighted Least Squares

Local Polynomial regression corresponds to locally polynomial estimator obtained from (locally) weighted least squares

$$\text{i.e. set } f(X_i) = \beta_0 + \beta_1(X_i - X) + \frac{\beta_2}{2!}(X_i - X)^2 + \dots + \frac{\beta_p}{p!}(X_i - X)^p$$

(local polynomial of degree p around X)

What you should know

Linear Regression

- Least Squares Estimator

- Normal Equations

- Gradient Descent

- Probabilistic Interpretation (connection to MCLE)

Regularized Linear Regression (connection to MCAP)

- Ridge Regression, Lasso

Beyond Linear

- Polynomial regression, Regression with Non-linear features, Bias-variance tradeoff, Kernelized ridge regression, Local Kernel Regression and Weighted Least Squares