Midwest Big Data Summer School: Machine Learning I: Introduction

Kris De Brabanter

kbrabant@iastate.edu

Iowa State University
Department of Statistics
Department of Computer Science

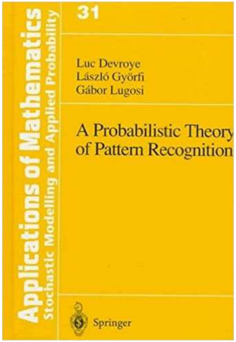
June 24, 2016

Outline

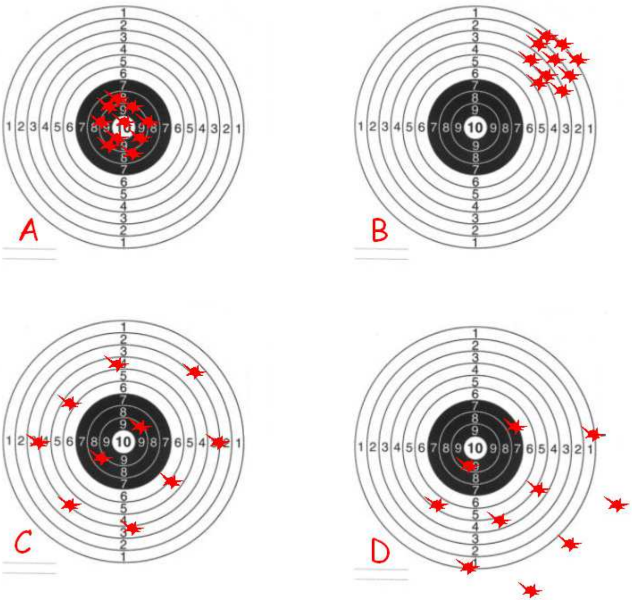
What is Pattern Recognition?

Devroye, Györfy & Lugosi (1996):

"Pattern recognition or discrimination is about guessing or predicting the unknown nature of an observation, a discrete quantity such as black or white, one or zero, sick or healthy, real or fake."



Two important concepts: Bias and Variance



Bayes' rule: An intuitive classifier

Key idea: Assign an observation x to class k such that $\mathbf{P}[Y = k \mid X = x]$ is largest.

Bayes' rule: An intuitive classifier

Key idea: Assign an observation x to class k such that $\mathbf{P}[Y = k \mid X = x]$ is largest.

Problem I: How to determine $\mathbf{P}[Y = k \mid X = x]$?

Bayes' rule: An intuitive classifier

Key idea: Assign an observation x to class k such that $\mathbf{P}[Y = k \mid X = x]$ is largest.

Problem I: How to determine $\mathbf{P}[Y = k \mid X = x]$?

This probability is usually not known/not readily available in practice.

Bayes' rule: An intuitive classifier

Key idea: Assign an observation x to class k such that $\mathbf{P}[Y = k \mid X = x]$ is largest.

Problem I: How to determine $\mathbf{P}[Y = k \mid X = x]$?

This probability is usually not known/not readily available in practice.

Bayes' rule (theorem):

$$\begin{aligned}
\underbrace{\mathbf{P}[Y = k \mid X = x]}_{\text{posterior probability}} &= \frac{\mathbf{P}[Y = k]\,\mathbf{P}[X = x \mid Y = k]}{\mathbf{P}[X = x]} \\
&= \frac{\mathbf{P}[Y = k]\,\mathbf{P}[X = x \mid Y = k]}{\sum_{i=1}^{K} \mathbf{P}[Y = i]\,\mathbf{P}[X = x \mid Y = i]}
\end{aligned}$$

Bayes' rule: An intuitive classifier

Key idea: Assign an observation x to class k such that $\mathbf{P}[Y = k \mid X = x]$ is largest.

Problem I: How to determine $\mathbf{P}[Y = k \mid X = x]$?

This probability is usually not known/not readily available in practice.

Bayes' rule (theorem):

$$\begin{aligned}\underbrace{\mathbf{P}[Y = k \mid X = x]}_{\text{posterior probability}} &= \frac{\mathbf{P}[Y = k]\,\mathbf{P}[X = x \mid Y = k]}{\mathbf{P}[X = x]} \\ &= \frac{\mathbf{P}[Y = k]\,\mathbf{P}[X = x \mid Y = k]}{\sum_{i=1}^{K} \mathbf{P}[Y = i]\,\mathbf{P}[X = x \mid Y = i]}\end{aligned}$$

or for continuous random variables X

$$\mathbf{P}[Y = k \mid X = x] = \frac{\mathbf{P}[Y = k] f(X = x \mid Y = k)}{\sum_{i=1}^{K} \mathbf{P}[Y = i] f(X = x \mid Y = i)}$$

Bayes' rule: An intuitive classifier

Since $\log$ is a strictly monotonic function and the denominator does not depend on k, assign class k s.t. the posterior probability is maximized

Bayes' rule: An intuitive classifier

Since $\log$ is a strictly monotonic function and the denominator does not depend on k, assign class k s.t. the posterior probability is maximized

$$
\begin{aligned}
\hat{Y}(x) &= \arg\max_k \log \mathbf{P}[Y = k \mid X = x] \\
&= \arg\max_k \log \left[\mathbf{P}[Y = k] f(X = x \mid Y = k)\right] \\
&= \arg\max_k \left[\log \mathbf{P}[Y = k] + \log f(X = x \mid Y = k)\right]
\end{aligned}
$$

Bayes' rule: An intuitive classifier

Since $\log$ is a strictly monotonic function and the denominator does not depend on k, assign class k s.t. the posterior probability is maximized

$$
\begin{aligned}
\hat{Y}(x) &= \arg\max_k \log \mathbf{P}[Y = k \mid X = x] \\
&= \arg\max_k \log \left[\mathbf{P}[Y = k] f(X = x \mid Y = k)\right] \\
&= \arg\max_k \left[\log \mathbf{P}[Y = k] + \log f(X = x \mid Y = k)\right]
\end{aligned}
$$

Problem II: Still 2 unknown quantities ($\mathbf{P}[Y = k]$ and $f(X = x \mid Y = k)$)

Bayes' rule: An intuitive classifier

Since $\log$ is a strictly monotonic function and the denominator does not depend on k, assign class k s.t. the posterior probability is maximized

$$
\begin{aligned}
\hat{Y}(x) &= \arg\max_k \log \mathbf{P}[Y = k \mid X = x] \\
&= \arg\max_k \log \left[\mathbf{P}[Y = k] f(X = x \mid Y = k)\right] \\
&= \arg\max_k \left[\log \mathbf{P}[Y = k] + \log f(X = x \mid Y = k)\right]
\end{aligned}
$$

Problem II: Still 2 unknown quantities ($\mathbf{P}[Y = k]$ and $f(X = x \mid Y = k)$)

$$\mathbf{P}\widehat{[Y = k]} = \frac{n_k}{n}$$

and for $f(X = x \mid Y = k)$, assume a density (usually normal) or estimate using kernel density estimation (or histogram)

Bayes' rule: An intuitive classifier

Assume

- $f(X = x \mid Y = k)$ multivariate normal with equal variance-covariance matrix for each class $\rightarrow$ LDA

Bayes' rule: An intuitive classifier

Assume

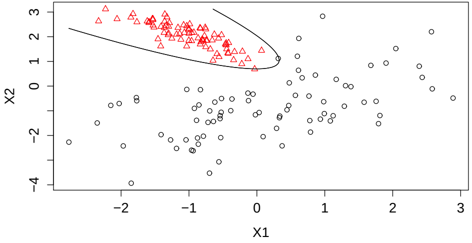
- $f(X = x \mid Y = k)$ multivariate normal with equal variance-covariance matrix for each class $\rightarrow$ LDA
- $f(X = x \mid Y = k)$ multivariate normal with unequal variance-covariance matrix for each class $\rightarrow$ QDA

Bayes' rule: An intuitive classifier

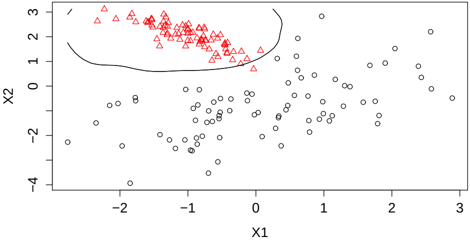
Assume

- $f(X = x \mid Y = k)$ multivariate normal with equal variance-covariance matrix for each class $\rightarrow$ LDA
- $f(X = x \mid Y = k)$ multivariate normal with unequal variance-covariance matrix for each class $\rightarrow$ QDA
- feature X_i is conditionally independent of every other feature X_j for $i \neq j$ given class k $\rightarrow$ Naive Bayes Classifier

Example: QDA vs. naive Bayes



(a) QDA



(b) Naive Bayes

k-Nearest Neighbor Classifier

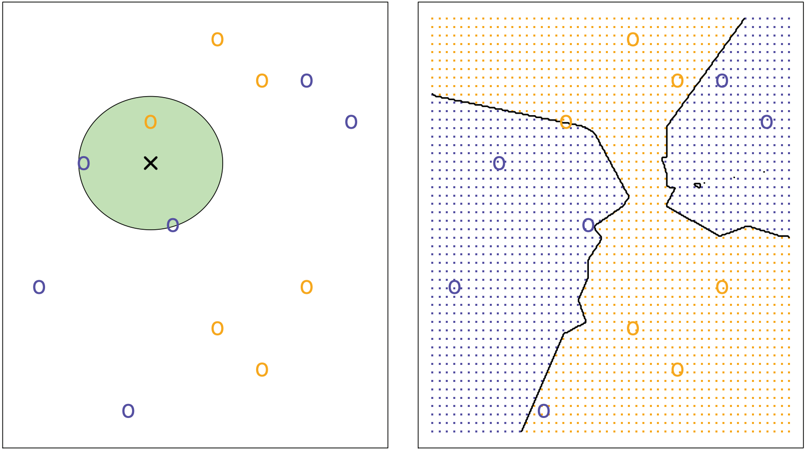


Figure: Illustration of k-Nearest Neighbor Classifier. Taken from Gareth, Witten, Hastie & Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer, 2013

Effect of k

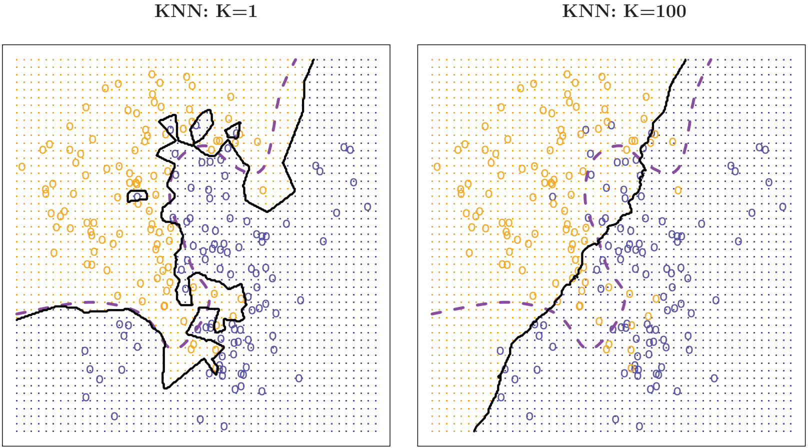


Figure: Effect of k in the Nearest Neighbor Classifier. Taken from Gareth, Witten, Hastie & Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer, 2013

Effect of k

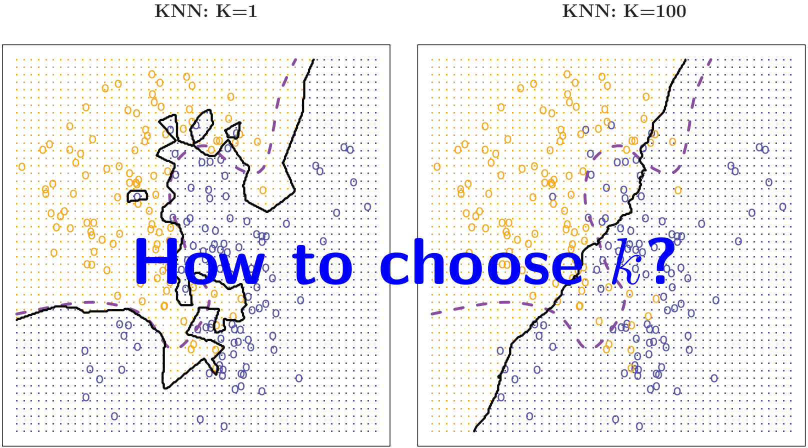


Figure: Effect of k in the Nearest Neighbor Classifier. Taken from Gareth, Witten, Hastie & Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer, 2013

Cross-validation

Leave-one-out Cross-validation

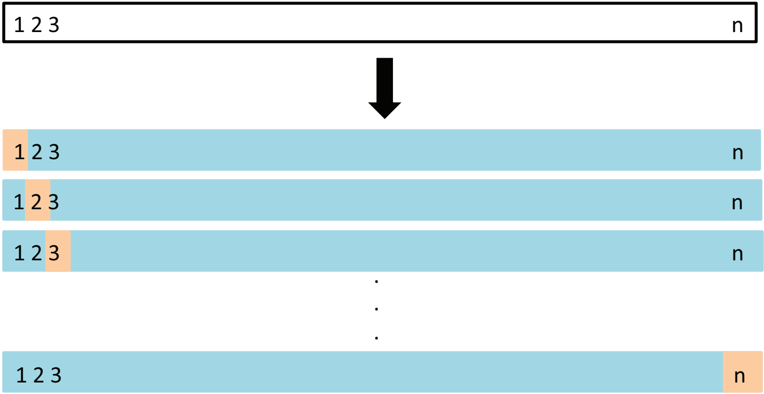
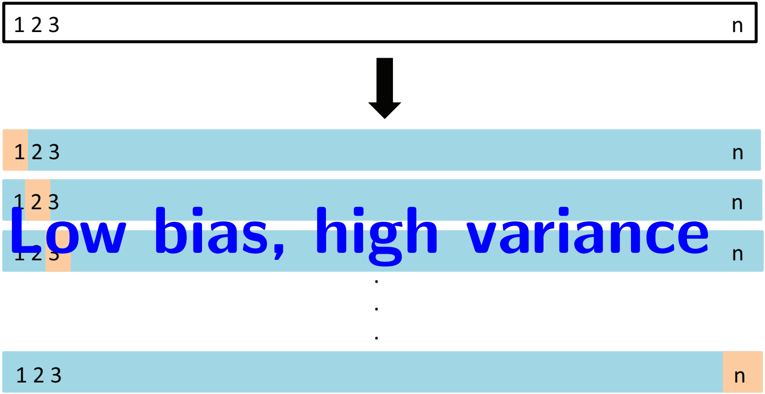


Figure: Leave-one-out Cross-Validation idea. Taken from Gareth, Witten, Hastie & Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer, 2013

Leave-one-out Cross-validation



Low bias, high variance

Figure: Leave-one-out Cross-Validation idea. Taken from Gareth, Witten, Hastie & Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer, 2013

v-fold Cross-Validation

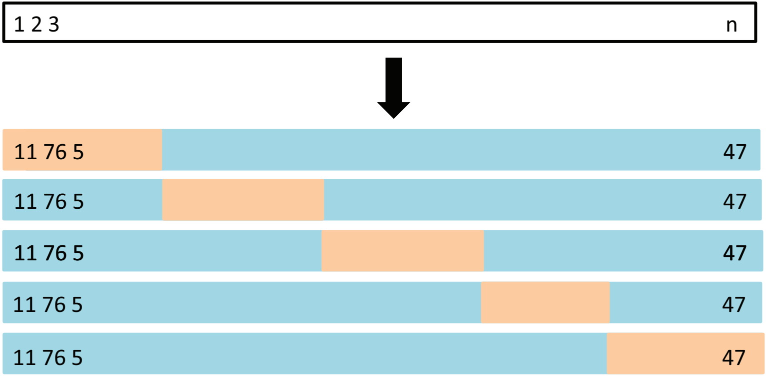


Figure: v-fold Cross-Validation idea. Taken from Gareth, Witten, Hastie & Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer, 2013

v-fold Cross-Validation

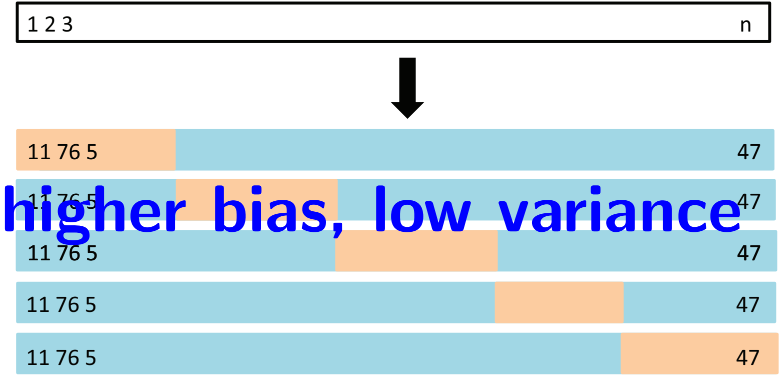


Figure: v-fold Cross-Validation idea. Taken from Gareth, Witten, Hastie & Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer, 2013

Cross-validation

$\boxed{\textbf{Cross-validation gives an ESTIMATE of the prediction error}}$

Cross-validation

Cross-validation gives an ESTIMATE of the prediction error

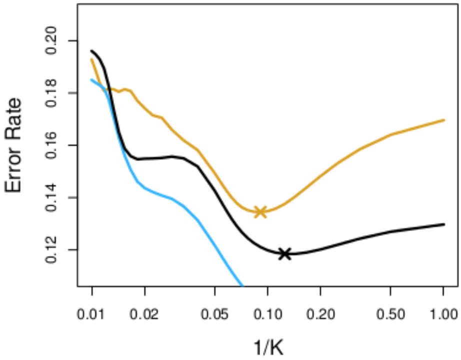


Figure: 10-fold CV error (black), test error (brown) and training error (blue). Taken from Gareth, Witten, Hastie & Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer, 2013

Drawbacks & other methods

- Easy to implement
- Can be applied to many situations
- Data-driven
- No explicit variance estimate of the errors needed

Drawbacks & other methods

- Easy to implement
- Can be applied to many situations
- Data-driven
- No explicit variance estimate of the errors needed
- Computationally expensive
- slow convergence

Drawbacks & other methods

- Easy to implement
- Can be applied to many situations
- Data-driven
- No explicit variance estimate of the errors needed
- Computationally expensive
- slow convergence

There exist other methods e.g. complexity criteria (AIC, BIC, Mallow's $C_p, \ldots$), generalized CV, etc.

Shrinkage methods

When $p > n$, OLS does not have a unique solution. Remember OLS

$$(\hat{\beta}_0, \dots, \hat{\beta}_p) = \underset{\beta_0,\dots,\beta_p \in \mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_i x_{ij} \right)^2 .$$

Shrinkage methods

When $p > n$, OLS does not have a unique solution. Remember OLS

$$(\hat{\beta}_0, \ldots, \hat{\beta}_p) = \underset{\beta_0,\ldots,\beta_p \in \mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_i x_{ij} \right)^2.$$

In matrix form: $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$

$$\mathbf{y} = (Y_1, \ldots, Y_n)^T, \beta = (\beta_0, \ldots, \beta_p)^T$$

and

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

Shrinkage methods

When $p > n$, OLS does not have a unique solution. Remember OLS

$$(\hat{\beta}_0, \ldots, \hat{\beta}_p) = \underset{\beta_0, \ldots, \beta_p \in \mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_i x_{ij} \right)^2 .$$

In matrix form: $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$

$$\mathbf{y} = (Y_1, \ldots, Y_n)^T, \beta = (\beta_0, \ldots, \beta_p)^T$$

and

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

$\Rightarrow \mathbf{X}^T \mathbf{X}$ will be (close to) singular and hence not invertible

Ridge regression

Solution to previous problem: make matrix $\mathbf{X}^T\mathbf{X}$ invertible by adding a little noise on the main diagonal

Ridge regression

Solution to previous problem: make matrix $\mathbf{X}^T\mathbf{X}$ invertible by adding a little noise on the main diagonal

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda \boldsymbol{I}_p)^{-1}\mathbf{X}^T\mathbf{y}, \quad \lambda \geq 0.$$

Ridge regression

Solution to previous problem: make matrix $\mathbf{X}^T \mathbf{X}$ invertible by adding a little noise on the main diagonal

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = (\mathbf{X}^T \mathbf{X} + \lambda \boldsymbol{I}_p)^{-1} \mathbf{X}^T \mathbf{y}, \quad \lambda \geq 0.$$

This corresponds to solving the following penalized residual sums of squares

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = \underset{\beta_0, \ldots, \beta_p}{\arg\min} \left\{ \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}.$$

Ridge regression

Solution to previous problem: make matrix $\mathbf{X}^T \mathbf{X}$ invertible by adding a little noise on the main diagonal

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = (\mathbf{X}^T \mathbf{X} + \lambda \boldsymbol{I}_p)^{-1} \mathbf{X}^T \mathbf{y}, \quad \lambda \geq 0.$$

This corresponds to solving the following penalized residual sums of squares

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = \underset{\beta_0, \ldots, \beta_p}{\arg\min} \left\{ \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}.$$

- as $\lambda \to 0$, $\hat{\boldsymbol{\beta}}^{\text{ridge}} \to \hat{\boldsymbol{\beta}}^{\text{OLS}}$
- as $\lambda \to \infty$, $\hat{\boldsymbol{\beta}}^{\text{ridge}} \to 0$
- shrinks the column space of $\mathbf{X}$ having small variance the most
- λ depends on σ_e^2, p and $\boldsymbol{\beta}$

Ridge regression

Solution to previous problem: make matrix $\mathbf{X}^T\mathbf{X}$ invertible by adding a little noise on the main diagonal

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\boldsymbol{I}_p)^{-1}\mathbf{X}^T\mathbf{y}, \quad \lambda \geq 0.$$

This corresponds to solving the following penalized residual sums of squares

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = \underset{\beta_0,\ldots,\beta_p}{\arg\min}\left\{\sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{p}x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right\}.$$

- as $\lambda \to 0$, $\hat{\boldsymbol{\beta}}^{\text{ridge}} \to \hat{\boldsymbol{\beta}}^{\text{OLS}}$
- as $\lambda \to \infty$, $\hat{\boldsymbol{\beta}}^{\text{ridge}} \to 0$
- shrinks the column space of $\mathbf{X}$ having small variance the most
- λ depends on σ_e^2, p and $\boldsymbol{\beta}$

Find λ via cross-validation

Why Ridge regression?

Theorem (Existence Theorem (Hoerl & Kennard, 1970))

There always exist a λ s.t. that $TMSE(\hat{\boldsymbol{\beta}}^{ridge}) < TMSE(\hat{\boldsymbol{\beta}}^{OLS})$

Why Ridge regression?

Theorem (Existence Theorem (Hoerl & Kennard, 1970))

There always exist a λ s.t. that $TMSE(\hat{\beta}^{ridge}) < TMSE(\hat{\beta}^{OLS})$

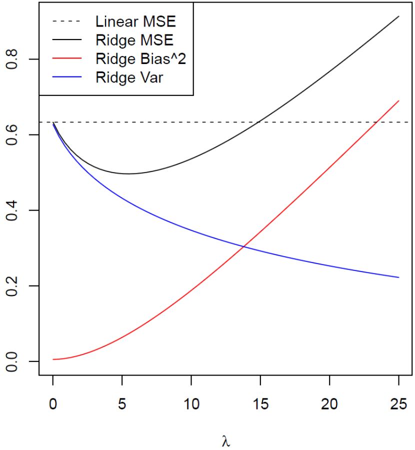
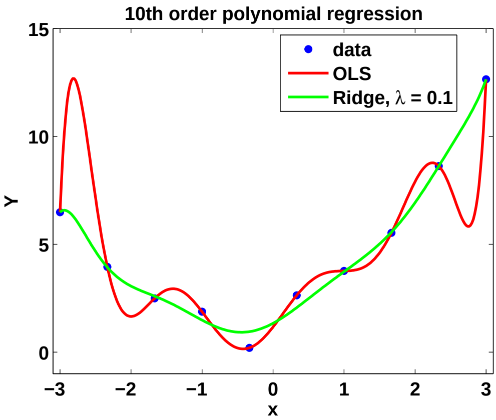


Figure: bias^Tbias and trace(variance) of ridge regression

The existence theorem in practice



OLS coeff.	Ridge coeff.
1.1660	1.3379
4.2225	1.7752
2.3333	1.5521
-5.3965	-0.9540
-0.8235	-0.3379
2.5254	0.3483
0.1503	0.0566
-0.4294	-0.0498
-0.0085	-0.0030
0.0235	0.0024

LASSO and variable selection (Tibshirani, 1996)

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \underset{\beta_0,\ldots,\beta_p}{\arg\min} \left\{ \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

LASSO and variable selection (Tibshirani, 1996)

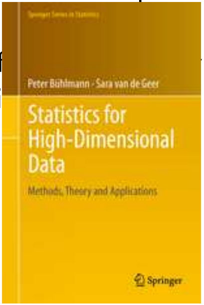
$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \underset{\beta_0,\ldots,\beta_p}{\arg\min} \left\{ \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

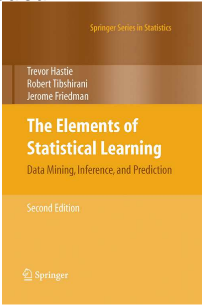
- No closed form, numerical optimization needed
- convex problem
- Sets certain coefficients β exactly equal to zero
- λ can be determined via cross-validation

LASSO and variable selection (Tibshirani, 1996)

$$\hat{\beta}^{\text{lasso}} = \underset{\beta_0,\dots,\beta_p}{\arg\min} \left\{ \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

- No closed form, numerical optimization needed
- convex problem
- Sets certain coeff … equal to
- λ can be determi … idation





References

- Bühlmann, P & van de Geer S. (2011), Statistics for High-Dimensional Data: Methods, Theory and Applications, Springer
- Devroye L., Györfi L. & Lugosi G. (1996). A Probabilistic Theory of Pattern Recognition, Springer
- Gareth J., Witten D., Hastie T. & Tibshirani R. (2013). An Introduction to Statistical Learning with Applications in R, Springer
- Hastie T., Tibshirani R. & Friedman J. (2009), The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd Ed.), Springer
- Hoerl A.E. & Kennard R. (1974), Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, 12: 55-67
- Györfi L, Kohler M., Krzyżak A. & Walk H. (2002). *A Distribution-Free Theory of Nonparametric Regression*, Springer
- Hampel F. R., Ronchetti E. M., Rousseeuw P. J. & Werner A. (1986), Robust Statistics: The Approach Based on Influence Functions, John Wiley & Sons
- Silverman B. W., Density Estimation for Statistics and Data Analysis, Chapman & Hall, 1986
- Simonoff J.S. (1996), *Smoothing Methods in Statistics*, Springer
- R. Tibshirani (1996), Regression shrinkage and selection via the lasso, *J. Royal. Statist. Soc B.*, 58(1): 267-288