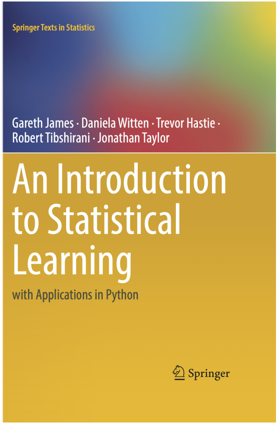
Making a Case

HN

February 18, 2025

Source of These Notes



Springer Texts in Statistics

Gareth James · Daniela Witten · Trevor Hastie · Robert Tibshirani · Jonathan Taylor

An Introduction to Statistical Learning

with Applications in Python

Springer

Notes from
An Introduction to Statistical Learning with Applications in Python

- It’s an easy read and as the name suggests, it’s just Introduction. Good for intuition building
- Its PDF is available for free from the Authors’ website.
- Sign-up & receive coupons (30%-50% off)

Last thing first (There is no magic wand)

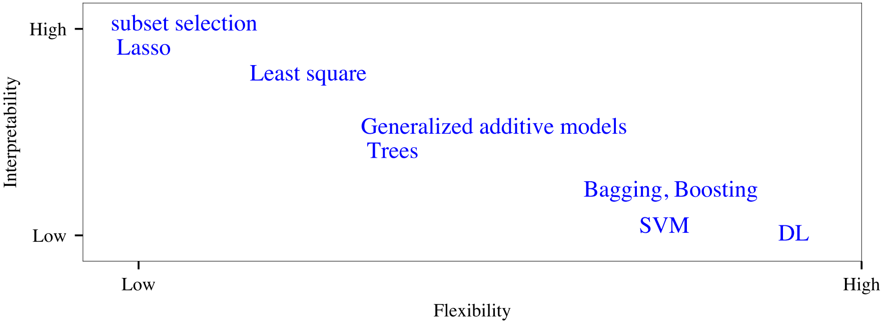
- There is no unique tool that outperforms other methods (there is no free lunch)

Last thing first (There is no magic wand)

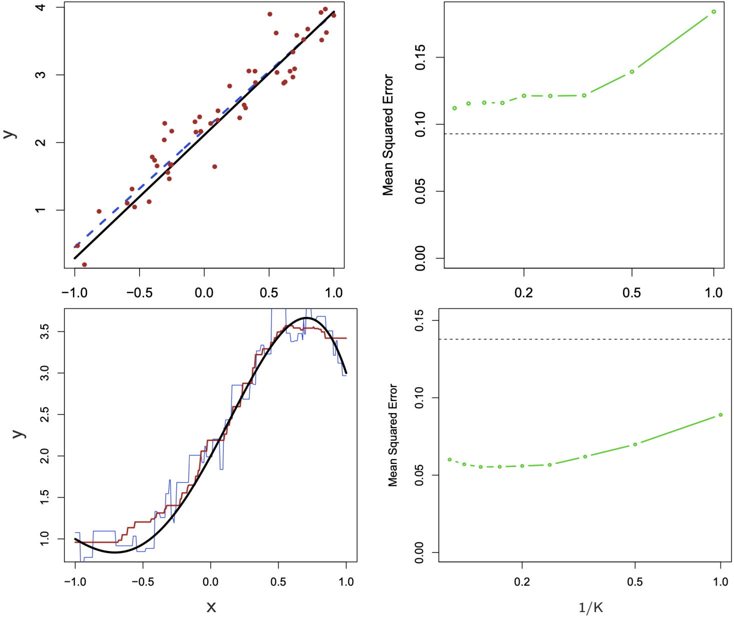
- There is no unique tool that outperforms other methods (there is no free lunch)
- There is no tool that is both accurate (in predicting) AND interpretable (unless we get lucky?!)

Last thing first (There is no magic wand)

- There is no unique tool that outperforms other methods (there is no free lunch)
- There is no tool that is both accurate (in predicting) AND interpretable (unless we get lucky?!)



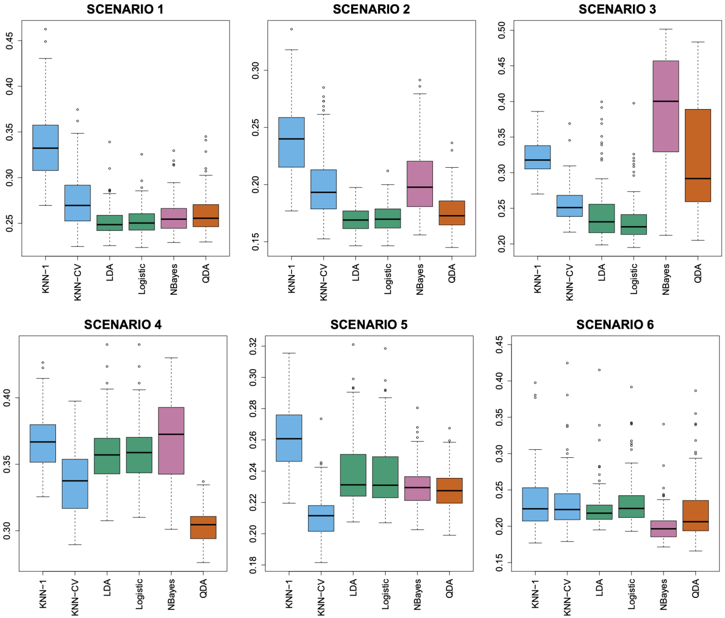
Comparison of Model Performances



Some Model Assumptions

- Logistic regression fits the data via likelihood maximization
- Linear Discriminant Analysis (LDA): distribution of X is normal in each class with identical covariance matrix.
- Quadratic Discriminant Analysis (QDA): distribution of X is normal in each class with class specific covariance matrix.

Comparison of Model Performances



PCR vs. PLS

Two methods for reducing the dimensionality in regression

- $y = \mathbf{X}\beta$

PCR vs. PLS

Two methods for reducing the dimensionality in regression

- $y = \mathbf{X}\beta$
- Principal Components Regression (PCR): regression using PCA

PCR vs. PLS

Two methods for reducing the dimensionality in regression

- $y = \mathbf{X}\beta$
- Principal Components Regression (PCR): regression using PCA
- Partial Least Squares (PLS)

PCR vs. PLS

Two methods for reducing the dimensionality in regression

- $y = \mathbf{X}\beta$
- Principal Components Regression (PCR): regression using PCA
- Partial Least Squares (PLS)
- "PLS is popular in the field of chemometrics … In practice it often performs no better than ridge regression or PCR. While the supervised dimension reduction of PLS can reduce bias, it also has the potential to increase variance, so that the overall benefit of PLS relative to PCR is a wash."

PCR vs. PLS

Two methods for reducing the dimensionality in regression

- $y = \mathbf{X}\beta$
- Principal Components Regression (PCR): regression using PCA
- Partial Least Squares (PLS)
- "PLS is popular in the field of chemometrics ... In practice it often performs no better than ridge regression or PCR. While the supervised dimension reduction of PLS can reduce bias, it also has the potential to increase variance, so that the overall benefit of PLS relative to PCR is a wash."

Other uses of PCA

1. Missing value imputation
2. Recommender systems
3. EDA

Some Notes on p-value

"..., p-values have recently been the topic of extensive commentary in the social science research community, to the extent that some social science journals have gone so far as to ban the use of p-values altogether! We will simply comment that when properly understood and applied, p-values provide a powerful tool for drawing inferential conclusions from our data."

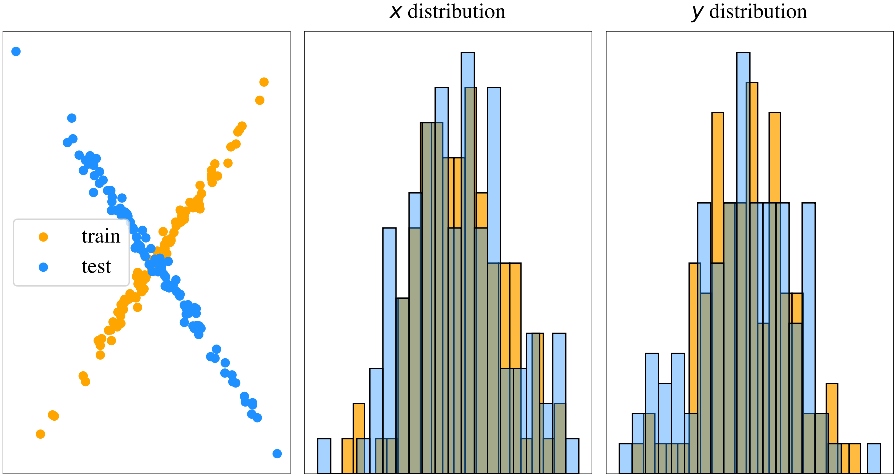
" ...By contrast, if we fail to reject H_0, then our findings are more nebulous: we will not know whether we failed to reject H_0 because our sample size was too small (in which case testing H_0 again on a larger or higher-quality dataset might lead to rejection), or whether we failed to reject H_0 because H_0 really holds." P. 559 (and read page 564).

Some Notes on p-value

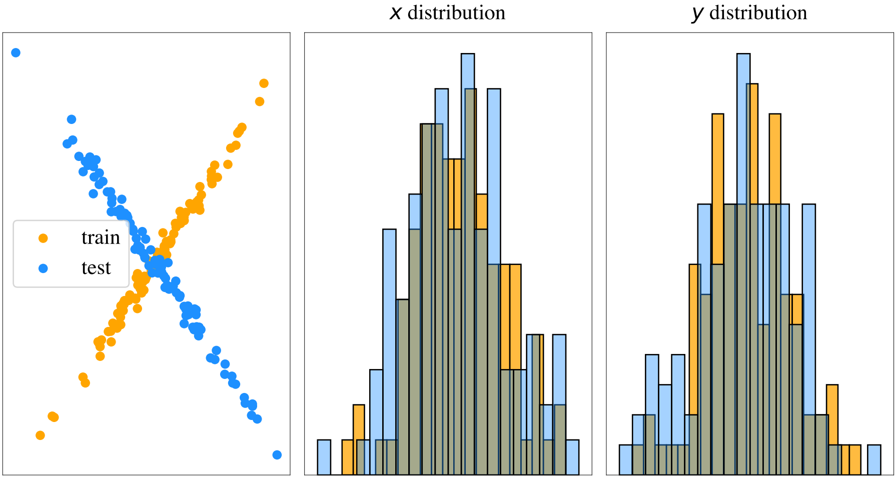
Remark

Hierarchical principle: *if we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant.*

Train Set v. Test Set

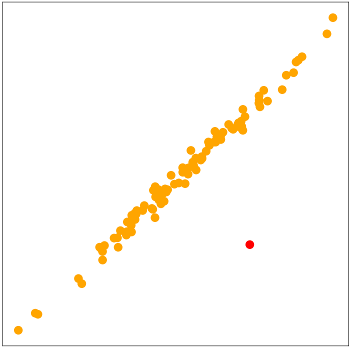


Train Set v. Test Set



- Earth Mover's Distance: 13
- Kullback Divergence: ∞

High-Leverage Points



We cannot look one dimension at a time to detect unusual points.

R^2

- True model $y = f(x) + \varepsilon$

R^2

- True model $y = f(x) + \varepsilon$
- Our model $y = \beta_0 + \beta_1 x + \varepsilon$

R^2

- True model $y = f(x) + \varepsilon$
- Our model $y = \beta_0 + \beta_1 x + \varepsilon$
- The error term ($\sim N(0, \sigma^2)$) is a catch-all for what we miss with this simple model: measurement errors, false modeling, omitted variables

R^2

- True model $y = f(x) + \varepsilon$
- Our model $y = \beta_0 + \beta_1 x + \varepsilon$
- The error term ($\sim N(0, \sigma^2)$) is a catch-all for what we miss with this simple model: measurement errors, false modeling, omitted variables
- RSS or RSE $= \sqrt{RSS/(n-p-1)}$ depend on Y-units.

R^2

- True model $y = f(x) + \varepsilon$
- Our model $y = \beta_0 + \beta_1 x + \varepsilon$
- The error term ($\sim N(0, \sigma^2)$) is a catch-all for what we miss with this simple model: measurement errors, false modeling, omitted variables
- RSS or RSE $= \sqrt{RSS/(n-p-1)}$ depend on Y-units.
- R^2 is unitless.

R^2

- True model $y = f(x) + \varepsilon$
- Our model $y = \beta_0 + \beta_1 x + \varepsilon$
- The error term ($\sim N(0, \sigma^2)$) is a catch-all for what we miss with this simple model: measurement errors, false modeling, omitted variables
- RSS or RSE $= \sqrt{RSS/(n-p-1)}$ depend on Y-units.
- R^2 is unitless.
- Small R^2 "might occur because the linear model is wrong, or the error variance σ^2 is high, or both."

R^2

- True model $y = f(x) + \varepsilon$
- Our model $y = \beta_0 + \beta_1 x + \varepsilon$
- The error term ($\sim N(0, \sigma^2)$) is a catch-all for what we miss with this simple model: measurement errors, false modeling, omitted variables
- RSS or RSE $= \sqrt{RSS/(n-p-1)}$ depend on Y-units.
- R^2 is unitless.
- Small R^2 "might occur because the linear model is wrong, or the error variance σ^2 is high, or both."
- "…and an R^2 value well below 0.1 might be more realistic!" (page 79)

Do I have enough data?

Do I have enough data? It depends. That question is ill imposed. Depends on the dimension (number of variables). So, it is about the ratio n/p. See figure below and page 22 of Hastie, Tibshirani, and Friedman, 2017 which is freely available too on Authors' website.

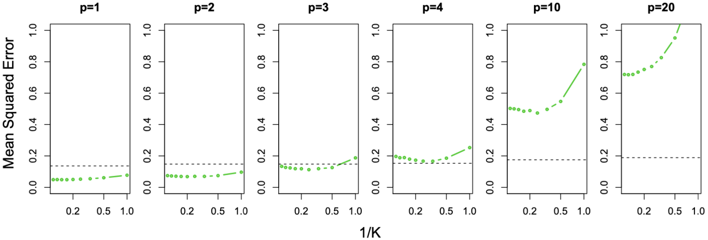


Figure: Test MSE for linear regression (black dashed lines) and KNN (green curves) as the number of variables p increases. The true function is non-linear in the first variable, as in the lower panel in Figure 3.19, and does not depend on the additional variables. The performance of linear regression deteriorates slowly in the presence of these additional noise variables, whereas KNN's performance degrades much more quickly as p increases.

General

- Anything worth making needs patience and practice
- If you want mastery, you need to immerse yourself in it till it becomes your second nature, just like walking

- Top(?) two take aways from graduate school
  1. Try not to have a bias/prejudice. be playful 🔍
  2. When you are wrong, you are wrong ⚠

More Resources

- First Course in Probability (Ross et al., 1976)
- Introduction to Linear Regression Analysis (Montgomery, Peck, and Vining, 2021)

Non-technical

- Guesstimation: (Weinstein and Adam, 2008)
- Excellent Sheep: The Miseducation of the American Elite and the Way to a Meaningful Life (Deresiewicz, 2014)
- The Culture Code: The Secrets of Highly Successful Groups (Coyle, 2018)

Tools

- Software Carpentry has lots of tutorials including GitHub!

Some Definition I

Definition (Outlier)

An outlier is a point for which y_i is far from the value predicted by the model

Definition (Outlier)

The training data contains outliers which are defined as observations that are far from the others.

There are methods to detect outliers and/or anomalies.

Definition (High leverage points)

unusual x value; "observations with high leverage have an unusual value for x_i."

Some Definition II

Definition (Variance of a method)

Variance refers to the amount by which $\hat{f}$ would change if we estimated it using a different training data set.

Definition (Bias of a method)

Bias refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model.

Bibliography I

- Coyle, D. (2018). *The Culture Code: The Secrets of Highly Successful Groups*. Bantam Books, an imprint of Random House, a division of Penguin Random House LLC New York.
- Deresiewicz, W. (2014). *Excellent Sheep: The Miseducation of the American Elite and the Way to a Meaningful Life*. Free Press.
- Hastie, T., Tibshirani, R., and Friedman, J. (2017). The elements of statistical learning: data mining, inference, and predi
- Montgomery, D. C., Peck, E. A., and Vining, G. G. (2021). *Introduction to Linear Regression Analysis*. John Wiley & Sons.
- Ross, S. M. et al. (1976). *A first course in probability*. Vol. 2. Macmillan New York.
- Weinstein, L. and Adam, J. A. (2008). *Guesstimation: Solving the world's problems on the back of a cocktail napkin*. Princeton University Press.