UNIVERSITÀ DI ROMA TOR VERGATA

EEBL - Statistical Learning

Revision - Week 3

1 Where to study

G James, D Witten, T Hastie, and R Tibshirani and J Friedman. *An Introduction to Statistical Learning with Applications in R.* Springer, Springer Series in Statistics.

- For assessing model accuracy and predictive performance and the bias-variance trade-off you should study section 2.2, pages 29-36,
- Cross-validation: see section 5.1, pages 198-206.
- Linear model selection, regularization and dimension reduction are dealt in chapter 6. Study pages 225-248, section 6.2.3. and sections 6.3.1 (Principal components regression).
- Principal components analysis is dealt with at more length in section 12.2.

You may want to browse what the other book has to say about those topics. This is not strictly needed, but *repetita iuvant*, as the saying goes. T Hastie, R Tibshirani and J Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition. Springer, Springer Series in Statistics, 2009.

Website: http://www-stat.stanford.edu/ElemStatLearn/.

- The linear regression model: chapter 3, sections 3.1, 3.2 up to page 51.
- Model evaluation and selection: chapter 7, sec. 7.1.–7.6. and 7.10. up to page 245.
- Subset selection: sections 3.3.1. and 3.3.2.
- Ridge regression: section 3.4.1.
- The Lasso: section 3.4.2.-3.4.3
- Principal Components Regression: section 3.5.1.

1.1 Addenda (useful things to know)

Let $\mathbf{X}$ denote a matrix of N standardized measurements on p variables, such that $\mathbf{X}'\mathbf{i} = \mathbf{0}$, i.e., the vector of means $\bar{\mathbf{x}} = \frac{1}{N}\mathbf{X}'\mathbf{i} = \mathbf{0}$, and $N^{-1}\sum_i x_{ik}^2 = 1$, $k = 1, \dots, p$.

Then $\mathbf{S} = \frac{1}{N}\mathbf{X}'\mathbf{X}$ is a correlation matrix. In fact, $s_{hk} = \frac{1}{N}\sum_{i=1}^N x_{ih}x_{ik}$ is the correlation coefficient of the h and k-th variables (the mean is zero and the variance is 1 for both variables). When the variables are not standardized, then $\mathbf{S} = \frac{1}{N}\mathbf{X}'\mathbf{X} - \bar{\mathbf{x}}\bar{\mathbf{x}}'$ is the covariance matrix of the x's,

Consider now the vector $\mathbf{a} = (a_1, \dots, a_p)'$. The vector $\mathbf{z} = \mathbf{X}\mathbf{a}$ contains the scores of the linear combination $z_i = a_1x_{i1} + a_2x_{i2} + \cdots + a_px_{ip}$. The mean of $\mathbf{z}$ is

$$\begin{aligned}\bar{z} &= \frac{1}{N}\mathbf{z}'\mathbf{i}\\ &= \frac{1}{N}\mathbf{a}'\mathbf{X}\mathbf{i}\\ &= \mathbf{a}'\bar{\mathbf{x}}\\ &= a_1\bar{x}_1 + a_2\bar{x}_2 + \cdots + a_p\bar{x}_p.\end{aligned}$$

If the variables are standardized, then $\bar{z} = 0$. The variance of $\mathbf{z}$ is

$$\begin{aligned}
s_z^2 &= \tfrac{1}{N}\sum(z_i - \bar{z})^2 \\
&= \tfrac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i^2 - \bar{z}^2 \\
&= \tfrac{1}{N}\mathbf{z}'\mathbf{z} - \bar{z}^2 \\
&= \tfrac{1}{N}\mathbf{a}'\mathbf{X}'\mathbf{X}\mathbf{a} - (\mathbf{a}'\bar{\mathbf{x}})^2 \\
&= \tfrac{1}{N}\mathbf{a}'\mathbf{X}'\mathbf{X}\mathbf{a} - \mathbf{a}'\bar{\mathbf{x}}\bar{\mathbf{x}}'\mathbf{a} \\
&= \mathbf{a}'\left(\tfrac{1}{N}\mathbf{X}'\mathbf{X} - \bar{\mathbf{x}}\bar{\mathbf{x}}'\right)\mathbf{a} \\
&= \mathbf{a}'\mathbf{S}\mathbf{a} \\
&= \sum_{j=1}^{p} a_j^2 s_j^2 + 2\sum_{j=1}^{p}\sum_{k=j+1}^{p} a_j a_k s_{jk}
\end{aligned}$$

The mean and variance of linear combinations are useful, e.g., in portfolio management.

1.2 Exercises and exam questions

The following provide you with a feeling of what a final exam question may look like.

1. *This question is worth 20 points (5+5+5+5)*
   During the course we considered the problem of predicting a quantitative outcome variable Y using the set of inputs $X = (X_1, \ldots, X_p)$. We modelled Y as follows: $Y = f(X) + \varepsilon$, where $f(X)$ is the systematic part (that can be predicted using the inputs X) and ε is a disturbance (error) term. We discussed in detail the linear regression model.

   (a) What are the main assumptions and characteristics of the specification of the linear regression model?

   (b) Suppose that $p = 1$ (simple regression) and that you have available a training sample $(y_i, x_i), i = 1, 2, \ldots, N$. Your regression model is $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. Illustrate the least squares learning method for estimating β_0 and β_1.

   (c) With reference to the previous point, let

   $$\frac{1}{N}\sum_{i=1}^{N} x_i = 4, \quad \frac{1}{N}\sum_{i=1}^{N} x_i^2 = 20, \quad \frac{1}{N}\sum_{i=1}^{N} y_i = 50, \quad \frac{1}{N}\sum_{i=1}^{N} x_i y_i = -80.$$

   Use the information above to compute the least squares estimate of the coefficients, $\hat{\beta}_1$ and $\hat{\beta}_0$.

   (d) The *expected training error* is the mean square error of prediction within the sample: $\mathrm{E}(\overline{err}|X) = \frac{1}{N}\sum_i b_i^2 + \sigma^2 - \sigma^2\frac{p+1}{N}$. It has a bias component (the first addend) and a component due to the variability of the predictions. Explain the fundamental bias-variance trade-off by saying what happens to the two components when we add or remove explanatory variables from the specification.

2. The following table summarizes the results of estimating a linear regression model for house prices from a training sample of $N = 1080$ observations:

```
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    4336.851  12084.856   0.359    0.720
sqft             97.862      3.485  28.082  < 2e-16 ***
Age            -694.671    141.292  -4.917 1.02e-06 ***
Pool            815.028   8926.466   0.091    0.927
Bedrooms     -20923.878   4567.168  -4.581 5.16e-06 ***
```

```
Fireplace      -97.130   5152.184  -0.019    0.985
Waterfront   63376.186   9389.874   6.749 2.43e-11 ***
DOM            -20.988     24.841  -0.845    0.398
---
Residual standard error: 76390 on 1072 degrees of freedom
Multiple R-squared: 0.6162,Adjusted R-squared: 0.6137
F-statistic: 245.9 on 7 and 1072 DF,  p-value: < 2.2e-16
```

- What is the interpretation of the p-value 0.40 associated to the explanatory variable `DOM` (days on the market)?
- What is the estimate of σ^2, the variance of the disturbance term in the regression model?
- Compute the value of the Bayesian Information Criterion, $BIC = \ln(RSS_p/N) + \frac{p+1}{N}\ln N$, using the information reported in the above summary table.
- What is the `F-statistic` in the last line meant for?
- Use a sentence to illustrate the role and the limitations of `Multiple R-squared` for assessing goodness of fit.

3. What is the meaning of the term “multicollinearity”? What are the consequences of multicollinearity?

4. Illustrate (in words) what happens when irrelevant variables are included as inputs in the regression model or relevant variables are omitted.

5. Consider the linear regression model for a training sample, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{y}$ is a vector of N observations on the output, and $\mathbf{X}$ is an $N \times (p+1)$ matrix of measurements on p inputs; the first column is a vector of 1's. The 'hat' matrix is defined as $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and the least squares residuals are $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$, where $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ are the fitted values. State which of the following statements are true:

   (a) The residuals have zero mean.

   (b) Ordinary least squares is always the best method for estimating $\boldsymbol{\beta}$.

   (c) The variance of the i-th training sample residual e_i is smaller than σ^2.

   (d) The residuals are uncorrelated with each of the p explanatory variables.

   (e) The residuals are uncorrelated with the variables not included in the model.

   (f) The fitted values have the same mean as the original observations y_i.

6. (Exam question) A large part of our course has been dedicated to model selection and evaluation.

   (a) What is meant by “bias-variance trade-off”?

   (b) Discuss the problem of measuring the predictive accuracy of a model, explaining the difference between the training error and the test error.

   (c) Present alternative ways of estimating the test error (out-of-sample predictive performance) (Mallow's C_p, cross-validation, etc.).

   (d) What is the main difference between AIC and BIC?

   (e) Discuss the pros and cons of the subset selection methodology known as forward stepwise.

7. *Exam question.*
In the linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{y}$ is a vector of N standardised observations on the output, and $\mathbf{X}$ is an $N \times p$ matrix of measurements on p standardised inputs, the ridge estimator of the regression coefficients is

$$\hat{\boldsymbol{\beta}}_r = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}.$$

   (a) Explain the rationale and the properties of this estimator, stressing the role of the shrinkage parameter λ.

   (b) Why do you need to standardise the inputs and the output?

   (c) What is LASSO, and how does it differ from ridge regression?

8. (Exam question) Illustrate cross-validation as a method for estimating the test error.

9. The table below presents the values of different model selection criteria for the different steps of a forward stepwise procedure applied to the model

```
model_full = price~sqft+I(sqft^2)+I(sqft^3)+Age+I(Age^2)+I(sqft*Age)+Pool+Baths+
             Bedrooms+Fireplace+Waterfront+DOM+factor(Occupancy)+factor(Style);
```

with the following specifications:

```
subr = regsubsets(model_full, data = train, method = "forward");
```

```
     Mallow's C_p  Adj R2    BIC        R2
 [1,] 207.652380 0.6813316 -674.3711 0.6818636
 [2,] 130.299229 0.7123160 -730.3515 0.7132765
 [3,]  84.474969 0.7308605 -764.9400 0.7322084
 [4,]  55.590869 0.7427162 -786.5808 0.7444343
 [5,]  32.510315 0.7522951 -803.9581 0.7543628
 [6,]  25.758862 0.7553768 -806.0836 0.7578271
 [7,]  19.982399 0.7580784 -807.3626 0.7609056
 [8,]  15.153086 0.7604092 -807.7886 0.7636090
 [9,]   9.874889 0.7629282 -808.7495 0.7664902
[10,]   6.288194 0.7647748 -808.0622 0.7687018
[11,]   5.490247 0.7655031 -804.5454 0.7698094
[12,]   5.650846 0.7658466 -800.0494 0.7705375
[13,]   6.461199 0.7659284 -794.8851 0.7710085
[14,]   7.438565 0.7659428 -789.5499 0.7714133
[15,]   8.751976 0.7658208 -783.8668 0.7716851
[16,]  10.382483 0.7655694 -777.8543 0.7718314
[17,]  12.282460 0.7652074 -771.5615 0.7718709
[18,]  14.178886 0.7648455 -765.2725 0.7719119
[19,]  16.102647 0.7644713 -758.9549 0.7719421
[20,]  18.034279 0.7640925 -752.6292 0.7719692
[21,]  20.001744 0.7636977 -746.2662 0.7719821
[22,]  22.000143 0.7632888 -739.8709 0.7719827
[23,]  24.000000 0.7628779 -733.4741 0.7719828
```

What step of the procedure is selected by each criterion? Why do you think that the R^2 does not a suitable criterion? How many explanatory variables are represented in the selected specifications?