Machine Learning: Basic Techniques

Alvaro J. Riascos Villegas
Universidad de los Andes and Quantil

July 6 2018

Contenido

Bayes regression and classification algorithms

- For the regression problem it is possible to prove that the best learning function, when the loss is cuadratic, is:

$$f(x) = E_P[Y \mid X]$$

- For the classification problem it is possible to prove that the best learning function, when the loss is zero - one, is: $f(x) = 1$ if $P(Y \mid X) \geq 0{,}5$ and zero otherwise. For multiple categories it is easily generalized.

Bayes regression and classification algorithms

- For the regression problem it is possible to prove that the best learning function, when the loss is cuadratic, is:

$$f(x) = E_P[Y \mid X]$$

- For the classification problem it is possible to prove that the best learning function, when the loss is zero - one, is: $f(x) = 1$ if $P(Y \mid X) \geq 0{,}5$ and zero otherwise. For multiple categories it is easily generalized.

Contenido

Linear Regression

- The Linear Regression model assumes:

$$f(x) \approx X^T \beta$$

- If we minimize the risk subject to the restriction that functions must be linear, we obtain:

$$\beta = E\left(XX^T\right)^{-1} E(XY)$$

- The model assumes that $f(x)$ is globally linear.

Linear Regression

- The Linear Regression model assumes:

$$f(x) \approx X^T \beta$$

- If we minimize the risk subject to the restriction that functions must be linear, we obtain:

$$\beta = E\left(XX^T\right)^{-1} E(XY)$$

- The model assumes that $f(x)$ is globally linear.

Linear Regression

- The Linear Regression model assumes:
$$f(x) \approx X^T \beta$$
- If we minimize the risk subject to the restriction that functions must be linear, we obtain:
$$\beta = E\left(XX^T\right)^{-1} E(XY)$$
- The model assumes that $f(x)$ is globally linear.

k-NN: k-Nearest Neighbors

- k-NN estimates the conditional expected value locally as a constant function.

$$f(x) \approx Mean\left(y \,|x \in N_k(x)\right)$$

Comparison between Linear Regression and k-NN

- Both methods approximate to $E(Y|X = x)$ using averages but they use different assumptions over the true learning function:
  - Linear Regression assumes that $f(x)$ is globally linear.
  - k-NN assumes that $f(x)$ is locally constant.

Comparison between Linear Regression and k-NN

- Both methods approximate to $E(Y|X = x)$ using averages but they use different assumptions over the true learning function:
  - Linear Regression assumes that $f(x)$ is globally linear.
  - k-NN assumes that $f(x)$ is locally constant.

Feature selection

- Two common problems:
  1. Test error: it's possible to diminish the test error by reducing the number of variables (this reduces the complexity and variance) although it increases bias.
  2. Interpretation: a smaller number of features allows an easier and better interpretation.
- We are going to discuss different ways of reducing the number of features.

Feature selection

- Two common problems:
  1. Test error: it's possible to diminish the test error by reducing the number of variables (this reduces the complexity and variance) although it increases bias.
  2. Interpretation: a smaller number of features allows an easier and better interpretation.
- We are going to discuss different ways of reducing the number of features.

Feature selection

- Two common problems:
  1. Test error: it's possible to diminish the test error by reducing the number of variables (this reduces the complexity and variance) although it increases bias.
  2. Interpretation: a smaller number of features allows an easier and better interpretation.
- We are going to discuss different ways of reducing the number of features.

Feature selection: Best subset of features

- Best subset of features.
  - The subset of features which produces the smallest test error is chosen.
  - Computationally expensive. Is computationally viable for models with less than 40 features.

Feature selection: Best subset of features

- Best subset of features.
  - The subset of features which produces the smallest test error is chosen.
  - Computationally expensive. Is computationally viable for models with less than 40 features.

Feature selection: Best subset of features

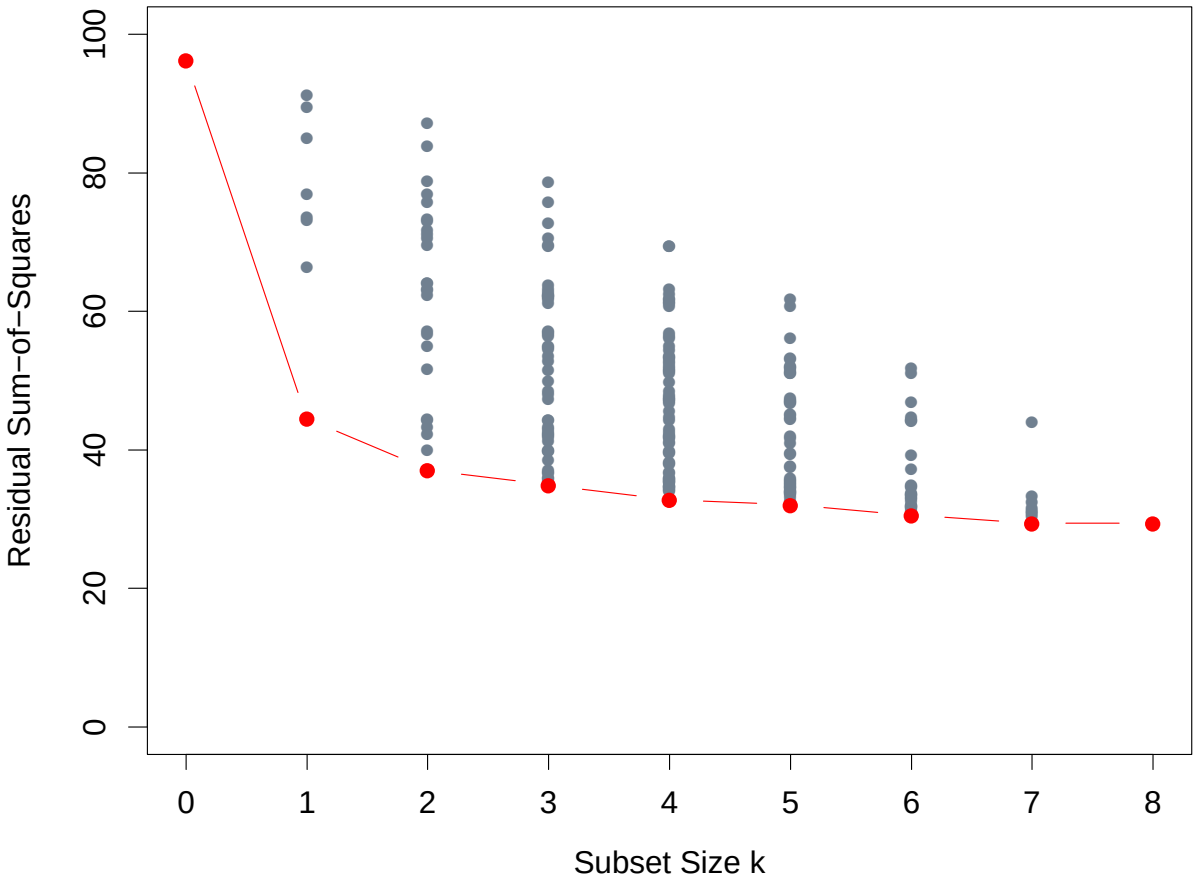


FIGURE 3.5. *All possible subset models for the prostate cancer example. At each subset size is shown the residual sum-of-squares for each model of that size.*

Feature selection: Forward, Backward and Stagewise selection

- Forward: Start with a model only with a constant and sequentially add the feature which reduces prediction error the most. On each stage the model is reestimated.
- Backward: Start with a model with all the features, proceed to eliminate the feature that less contributes to the final prediction (it can be done using the Z-score). On each stage the model is reestimated.
- Stagewise: Start with a model only with a constant and sequentially add the feature that has the biggest correlation with the residual errors of the last model. Don't reestimate.

Feature selection: Forward, Backward and Stagewise selection

- Forward: Start with a model only with a constant and sequentially add the feature which reduces prediction error the most. On each stage the model is reestimated.
- Backward: Start with a model with all the features, proceed to eliminate the feature that less contributes to the final prediction (it can be done using the Z-score). On each stage the model is reestimated.
- Stagewise: Start with a model only with a constant and sequentially add the feature that has the biggest correlation with the residual errors of the last model. Don't reestimate.

Feature selection: Forward, Backward and Stagewise selection

- Forward: Start with a model only with a constant and sequentially add the feature which reduces prediction error the most. On each stage the model is reestimated.
- Backward: Start with a model with all the features, proceed to eliminate the feature that less contributes to the final prediction (it can be done using the Z-score). On each stage the model is reestimated.
- Stagewise: Start with a model only with a constant and sequentially add the feature that has the biggest correlation with the residual errors of the last model. Don't reestimate.

Feature selection: Growth Regressions

Comparing Variable Selection Algorithms: Which Variables Appeared as Important Predictors of Economic Growth?

Predictor	*Bayesian model averaging*	*CDF(0)*	*LASSO*	*Spike-and-Slab*
GDP level 1960	1.000	1.000	-	0.9992
Fraction Confucian	0.995	1.000	2	0.9730
Life expectancy	0.946	0.942	-	0.9610
Equipment investment	0.757	0.997	1	0.9532
Sub-Saharan dummy	0.656	1.000	7	0.5834
Fraction Muslim	0.656	1.000	8	0.6590
Rule of law	0.516	1.000	-	0.4532
Open economy	0.502	1.000	6	0.5736
Degree of capitalism	0.471	0.987	9	0.4230
Fraction Protestant	0.461	0.966	5	0.3798

Source: The table is based on that in Ley and Steel (2009); the data analyzed is from Sala-i-Martín (1997).
Notes: We illustrate different methods of variable selection. This exercise involved examining a dataset of 72 counties and 42 variables in order to see which variables appeared to be important predictors of economic growth. The table shows ten predictors that were chosen by Sala-i-Martín (1997) using a CDF(0) measure defined in the 1997 paper; Ley and Steel (2009) using Bayesian model averaging, LASSO, and spike-and-slab regressions. Metrics used are not strictly comparable across the various models. The "Bayesian model averaging" and "Spike-and-Slab" columns are posterior probabilities of inclusion; the "LASSO" column just shows the ordinal importance of the variable or a dash indicating that it was not included in the chosen model; and the CDF(0) measure is defined in Sala-i-Martín (1997).

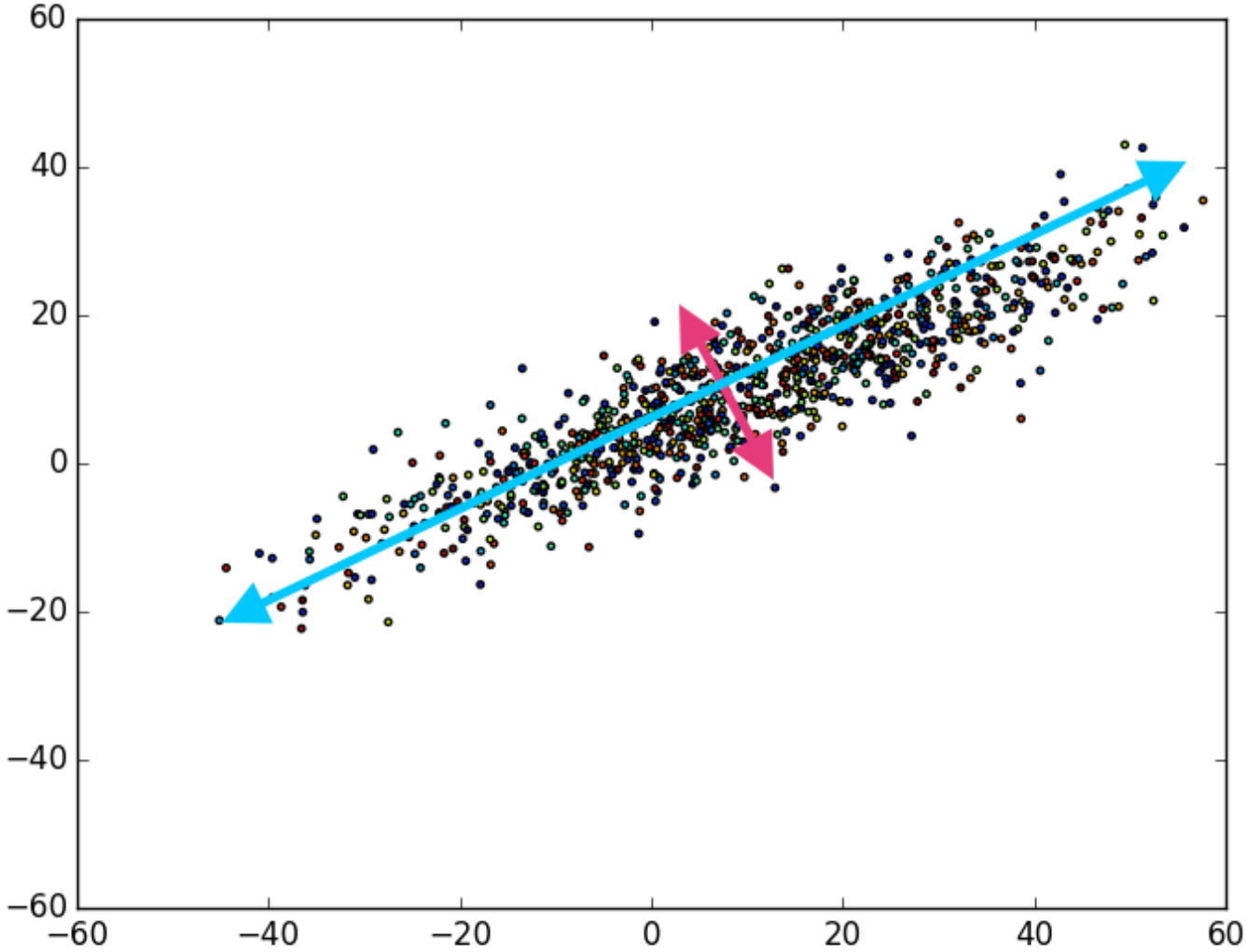
Principal components

- Sometimes many features are highly correlated.
- The information they give to the model is redundant.
- Can we construct a few new features that explain most of the variation that the original features contain?
- What if we wanted to build a single feature to replace all the others, how is this possible?

Principal components

- Sometimes many features are highly correlated.
- The information they give to the model is redundant.
- Can we construct a few new features that explain most of the variation that the original features contain?
- What if we wanted to build a single feature to replace all the others, how is this possible?

Principal components



Principal components

- The principal components are new features.
- They are linear combinations of the orignal features.
- The first principal component is the linear combination that maximizes the variance.
- The second principal component is the linear combination that maximizes the variance subject to being orthogonal to the first principal component.
- Successively, p principal components can be build.

Principal components

- The principal components are new features.
- They are linear combinations of the orignal features.
- The first principal component is the linear combination that maximizes the variance.
- The second principal component is the linear combination that maximizes the variance subject to being orthogonal to the first principal component.
- Successively, p principal components can be build.

Principal components

- The principal components are new features.
- They are linear combinations of the orignal features.
- The first principal component is the linear combination that maximizes the variance.
- The second principal component is the linear combination that maximizes the variance subject to being orthogonal to the first principal component.
- Successively, p principal components can be build.

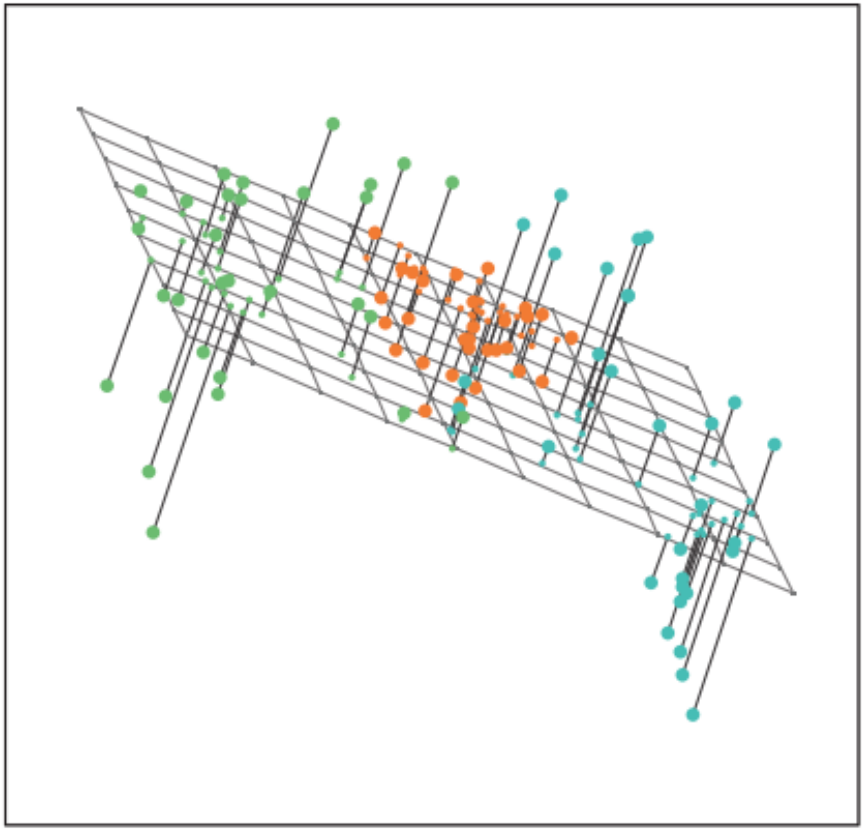
Principal components

- The principal components are new features.
- They are linear combinations of the orignal features.
- The first principal component is the linear combination that maximizes the variance.
- The second principal component is the linear combination that maximizes the variance subject to being orthogonal to the first principal component.
- Successively, p principal components can be build.

Principal components

- The principal components are new features.
- They are linear combinations of the orignal features.
- The first principal component is the linear combination that maximizes the variance.
- The second principal component is the linear combination that maximizes the variance subject to being orthogonal to the first principal component.
- Successively, p principal components can be build.

Principal Components



Aplication: handwritten numbers

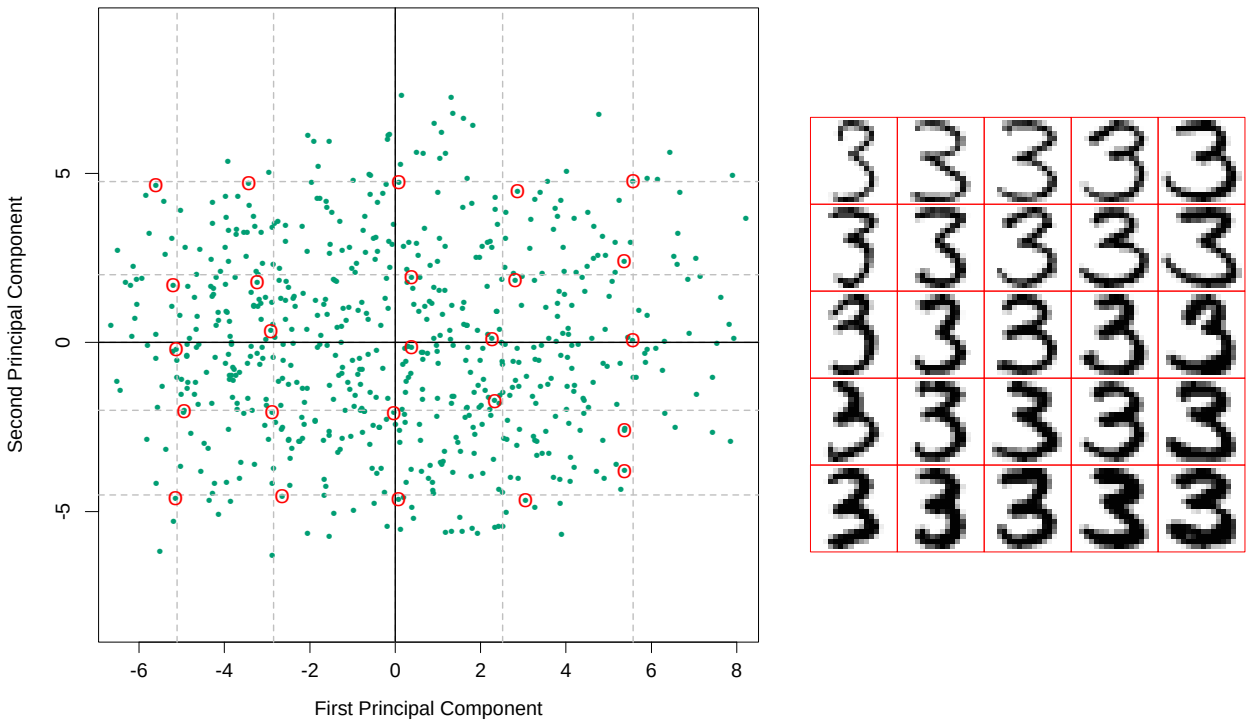


FIGURE 14.23. *(Left panel:) the first two principal components of the handwritten threes. The circled points are the closest projected images to the vertices of a grid, defined by the marginal quantiles of the principal components. (Right panel:) The images corresponding to the circled points. These show the nature of the first two principal components.*

PCA in practice

1. Apply PCA.
2. Choose the first PCs that explain the x % of the data's variance.
3. Make a classification model using the PCs.

- This can make a 200 feature model into a 5 feature model.

Contenido

Regularization

- The best subset can have the smallest test error. But it's a discrete method. Features are either included or excluded fully.
- This makes the method to have a high variance.
- Regularization methods are more continuous and have smaller variance.
- Lets consider Ridge Regression and Lasso.
- **Regularization is key concept in machine learning**. A must for economists.
- It allows to control for complexity, trading bias for variance.

Regularization

- The best subset can have the smallest test error. But it's a discrete method. Features are either included or excluded fully.
- This makes the method to have a high variance.
- Regularization methods are more continuous and have smaller variance.
- Lets consider Ridge Regression and Lasso.
- **Regularization is key concept in machine learning**. A must for economists.
- It allows to control for complexity, trading bias for variance.

Regularization

- The best subset can have the smallest test error. But it's a discrete method. Features are either included or excluded fully.
- This makes the method to have a high variance.
- Regularization methods are more continuous and have smaller variance.
- Lets consider Ridge Regression and Lasso.
- **Regularization is key concept in machine learning**. A must for economists.
- It allows to control for complexity, trading bias for variance.

Regularization

- The best subset can have the smallest test error. But it's a discrete method. Features are either included or excluded fully.
- This makes the method to have a high variance.
- Regularization methods are more continuous and have smaller variance.
- Lets consider Ridge Regression and Lasso.
- **Regularization is key concept in machine learning**. A must for economists.
- It allows to control for complexity, trading bias for variance.

Regularization

- The best subset can have the smallest test error. But it's a discrete method. Features are either included or excluded fully.
- This makes the method to have a high variance.
- Regularization methods are more continuous and have smaller variance.
- Lets consider Ridge Regression and Lasso.
- **Regularization is key concept in machine learning**. A must for economists.
- It allows to control for complexity, trading bias for variance.

Regularization

- The best subset can have the smallest test error. But it's a discrete method. Features are either included or excluded fully.
- This makes the method to have a high variance.
- Regularization methods are more continuous and have smaller variance.
- Lets consider Ridge Regression and Lasso.
- **Regularization is key concept in machine learning**. A must for economists.
- It allows to control for complexity, trading bias for variance.

Regularization: Ridge

- Solve:

$$\min_{\beta_0,\beta}\{\sum_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta)^2 + \lambda\|(\beta)\|^2\} \quad (1)$$

Regularization: Ridge

TABLE 3.3. *Estimated coefficients and test error results, for different subset and shrinkage methods applied to the prostate data. The blank entries correspond to variables omitted.*

Term	LS	Best Subset	Ridge	Lasso	PCR	PLS
`Intercept`	2.465	2.477	2.452	2.468	2.497	2.452
`lcavol`	0.680	0.740	0.420	0.533	0.543	0.419
`lweight`	0.263	0.316	0.238	0.169	0.289	0.344
`age`	−0.141		−0.046		−0.152	−0.026
`lbph`	0.210		0.162	0.002	0.214	0.220
`svi`	0.305		0.227	0.094	0.315	0.243
`lcp`	−0.288		0.000		−0.051	0.079
`gleason`	−0.021		0.040		0.232	0.011
`pgg45`	0.267		0.133		−0.056	0.084
Test Error	0.521	0.492	0.492	0.479	0.449	0.528
Std Error	0.179	0.143	0.165	0.164	0.105	0.152

Regularization: Ridge

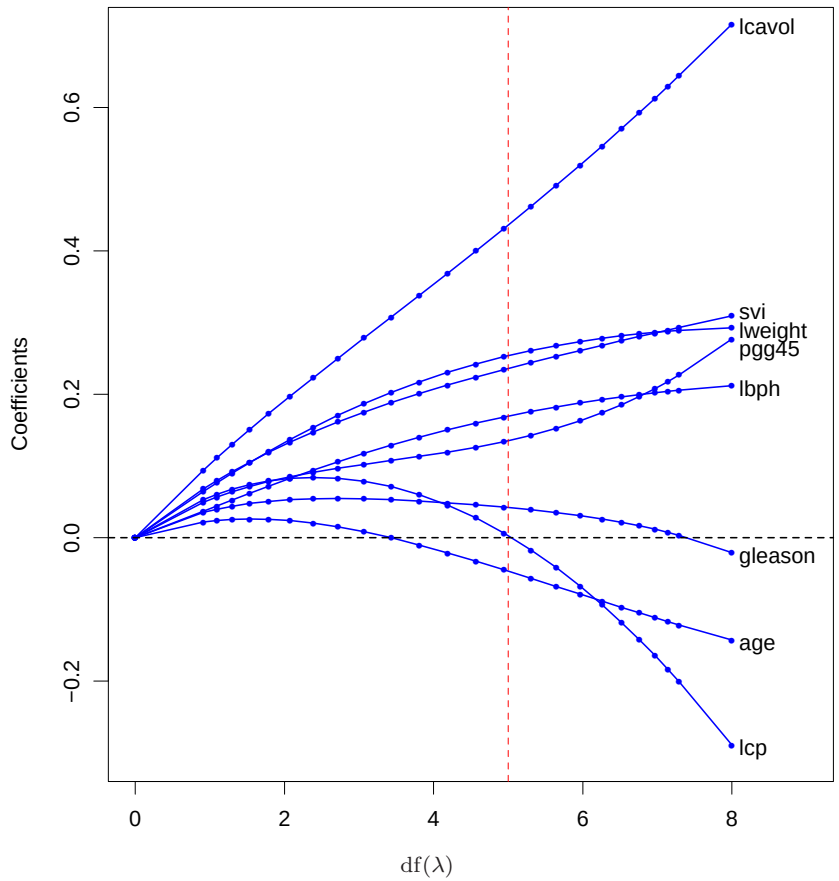


FIGURE 3.8. *Profiles of ridge coefficients for the prostate cancer example, as the tuning parameter* λ *is varied. Coefficients are plotted versus* $\mathrm{df}(\lambda)$*, the effective degrees of freedom. A vertical line is drawn at* $\mathrm{df} = 5.0$*, the value chosen by cross-validation.*

Summing up

- There are several techniques for controlling models complexity.
- In the k-NN model this is controlled with k.
- In the linear regression model this is controlled with: variables selection techniques (subset selection, backward, forward, PCA, etc.) and regularization.
- Controlling complexity trades bias for variance, looking forward to reduce test error.

Summing up

- There are several techniques for controlling models complexity.
- In the k-NN model this is controlled with k.
- In the linear regression model this is controlled with: variables selection techniques (subset selection, backward, forward, PCA, etc.) and regularization.
- Controlling complexity trades bias for variance, looking forward to reduce test error.

Summing up

- There are several techniques for controlling models complexity.
- In the k-NN model this is controlled with k.
- In the linear regression model this is controlled with: variables selection techniques (subset selection, backward, forward, PCA, etc.) and regularization.
- Controlling complexity trades bias for variance, looking forward to reduce test error.

Summing up

- There are several techniques for controlling models complexity.
- In the k-NN model this is controlled with k.
- In the linear regression model this is controlled with: variables selection techniques (subset selection, backward, forward, PCA, etc.) and regularization.
- Controlling complexity trades bias for variance, looking forward to reduce test error.

Contenido

CART

- The best classifier can be expressed by:

$$f(x) = \sum_{m=1}^{M} c_m I\{x \in R_m\} \tag{2}$$

where R_m are different regions in where the function is approximated by a constant c_m.

- Some regions are difficult to describe (left top panel, next figure).
- An alternative is to find regions by making sequential binary partitions.

CART

- The best classifier can be expressed by:

$$f(x) = \sum_{m=1}^{M} c_m I\{x \in R_m\} \tag{2}$$

where R_m are different regions in where the function is approximated by a constant c_m.

- Some regions are difficult to describe (left top panel, next figure).
- An alternative is to find regions by making sequential binary partitions.

CART

- The best classifier can be expressed by:

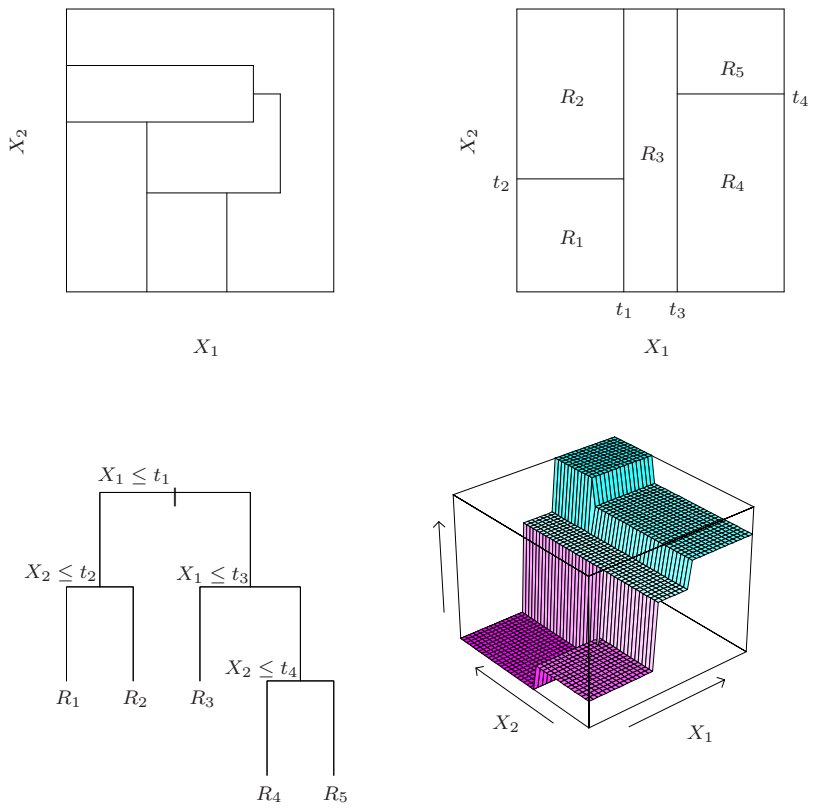
$$f(x) = \sum_{m=1}^{M} c_m I\{x \in R_m\} \tag{2}$$

where R_m are different regions in where the function is approximated by a constant c_m.

- Some regions are difficult to describe (left top panel, next figure).
- An alternative is to find regions by making sequential binary partitions.

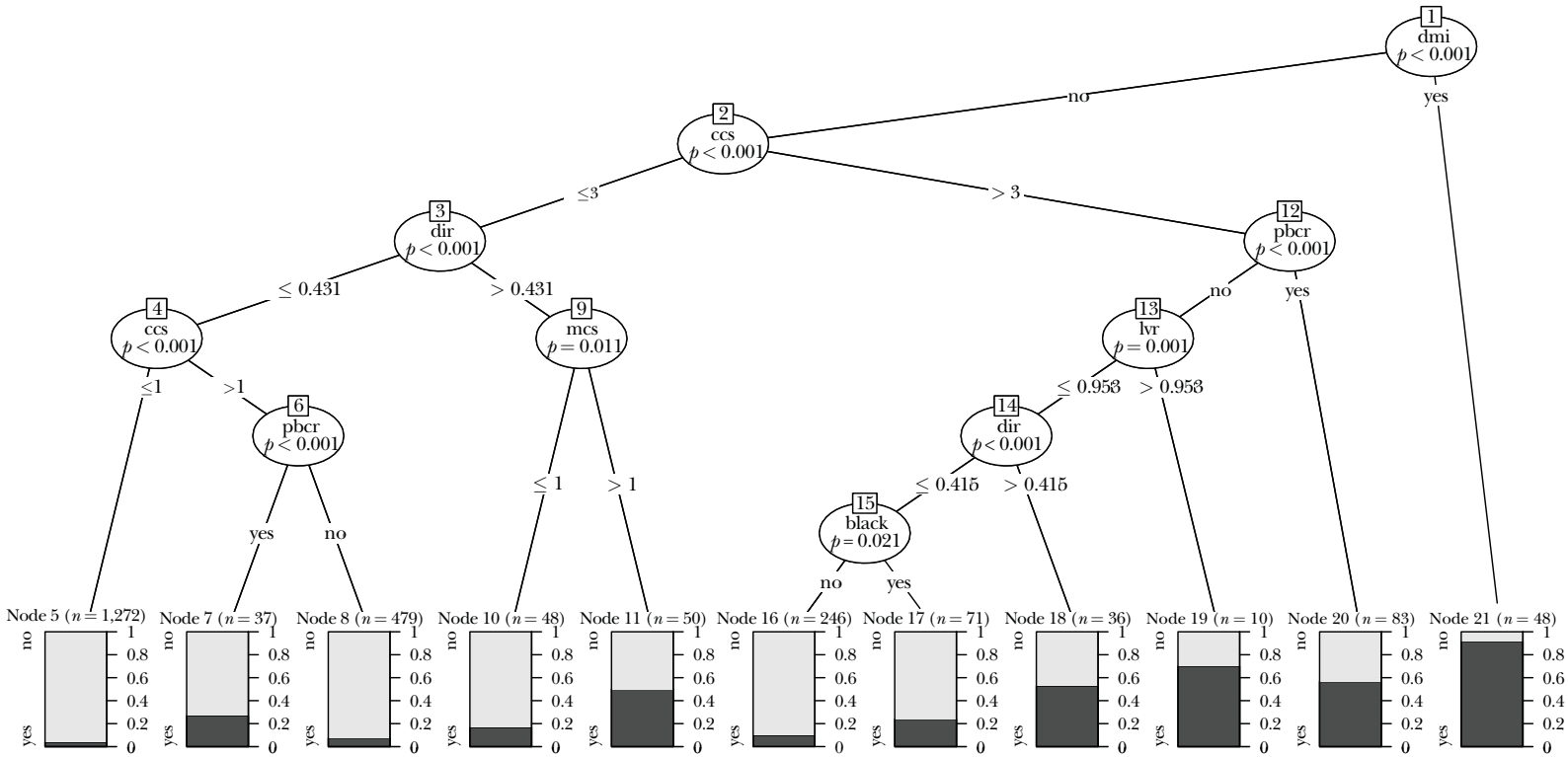
Classification and Regression Trees (CART)

- Panels 2, 3, 4 represent regions obtain by making sequential binary partitions.



Example: Home Mortgage Disclosure Act

Home Mortgage Disclosure Act (HMDA) Data Tree



Notes: Figure 5 shows a conditional tree estimated using the **R** package **party**. The black bars indicate the fraction of each group who were denied mortgages. The most important determinant of this is the variable "dmi," or "denied mortgage insurance." Other variables are: "dir," debt payments to total income ratio; "hir," housing expenses to income ratio; "lvr," ratio of size of loan to assessed value of property; "ccs," consumer credit score; "mcs," mortgage credit score; "pbcr," public bad credit record; "dmi," denied mortgage insurance; "self," self-employed; "single," applicant is single; "uria," 1989 Massachusetts unemployment rate applicant's industry; "condominium," unit is condominium; "black," race of applicant black; and "deny," mortgage application denied.

Classification Trees

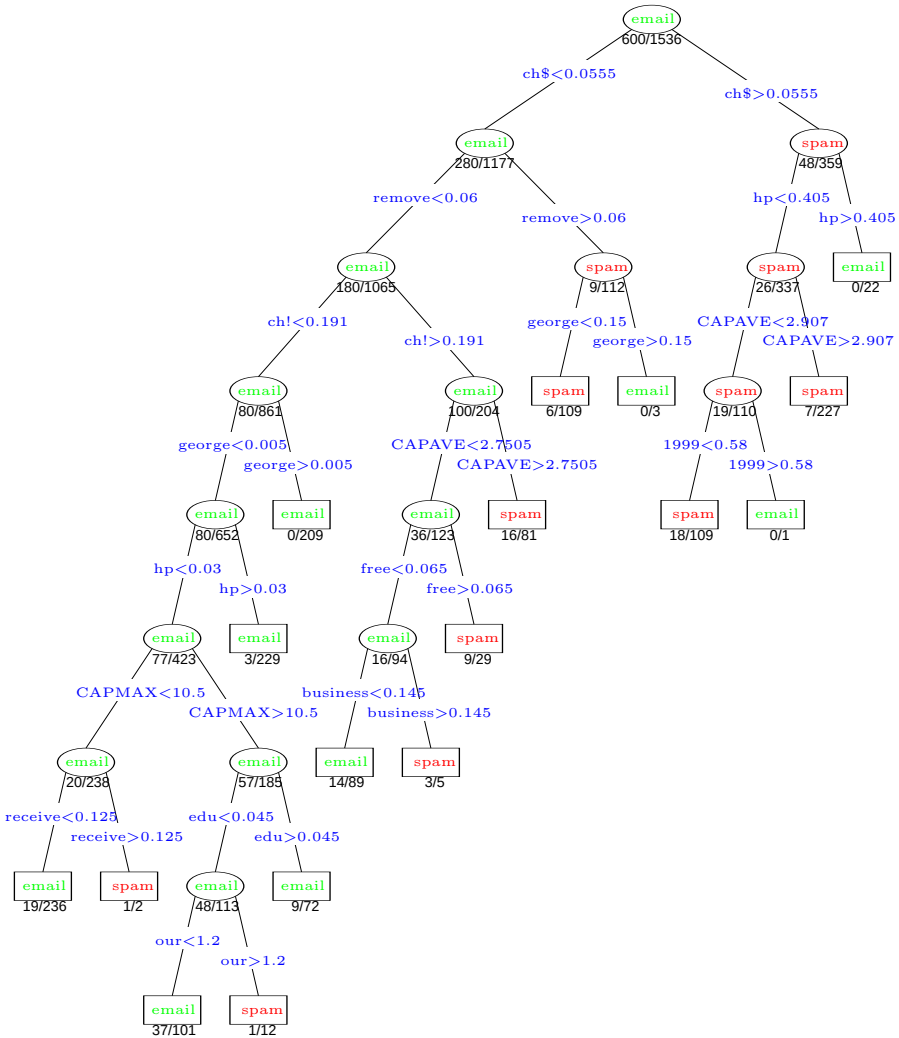


FIGURE 9.5. *The pruned tree for the* `spam` *example. The split variables are shown in blue on the branches, and the classification is shown in every node. The numbers under the terminal nodes indicate misclassification rates on the test data.*

Classification Trees

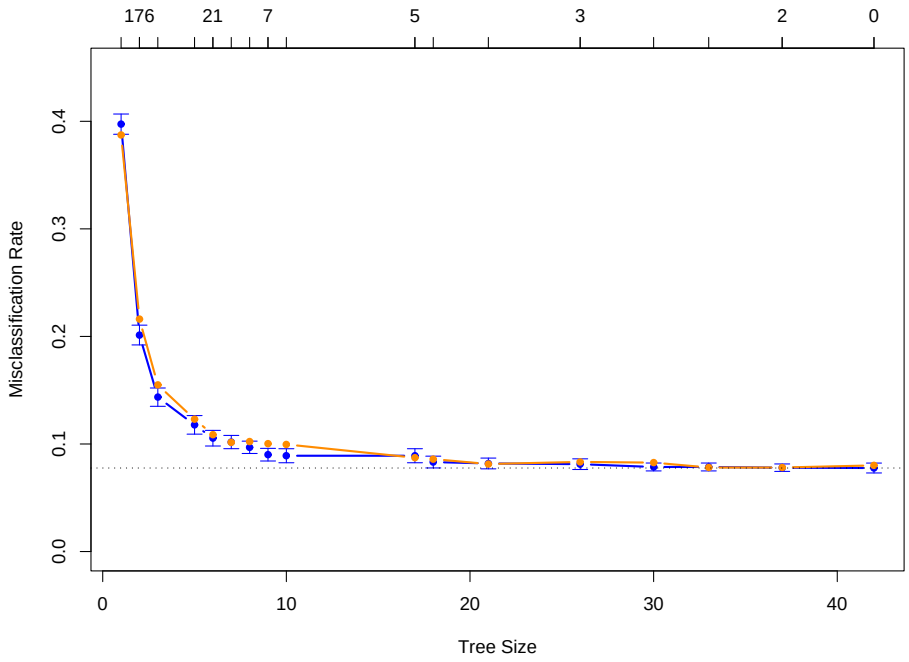


FIGURE 9.4. *Results for* `spam` *example. The blue curve is the* 10*-fold cross-validation estimate of misclassification rate as a function of tree size, with standard error bars. The minimum occurs at a tree size with about* 17 *terminal nodes (using the "one-standard--error" rule). The orange curve is the test error, which tracks the CV error quite closely. The cross-validation is indexed by values of* α*, shown above. The tree sizes shown below refer to* $|T_\alpha|$*, the size of the original tree indexed by* α*.*

Contenido

Random Forests

- It is a technique that consists of the construction of many uncorrelated trees and then averaging them. By doing that, the variance is reduced.

Random Forests: Algorithm

- For $b = 1, ..., B$
- Create samples of Z^* of the same size of the sample.
- Grow a tree T_b using the next steps until we reach a number of observations less than n_{min} in each node.
  1. Randomly, select m variables from the p variables.
  2. Choose the best partition of the m variables in the node.
  3. Repeat 1 and 2 until we reach the minimum in each leaf.
- In the regression task, average the trees. In the classification task, use majority vote (each tree contributes with one vote).

Random Forests: Properties

- When each tree grows sufficiently large it is possible to decrease the bias (increasing the variance). With the average, the variance is reduced.
- The average bias is similar to the bias of each tree because each tree is built using the same process. Hence, the potential benefits are found in the decrease of the variance.
- The variance reduction is obtained efficiently by averaging uncorrelated trees.
- The random choice in each partition guarantees that few variables don't dominate the regression.

Random Forests: Properties

- When each tree grows sufficiently large it is possible to decrease the bias (increasing the variance). With the average, the variance is reduced.
- The average bias is similar to the bias of each tree because each tree is built using the same process. Hence, the potential benefits are found in the decrease of the variance.
- The variance reduction is obtained efficiently by averaging uncorrelated trees.
- The random choice in each partition guarantees that few variables don't dominate the regression.

Random Forests: Properties

- When each tree grows sufficiently large it is possible to decrease the bias (increasing the variance). With the average, the variance is reduced.
- The average bias is similar to the bias of each tree because each tree is built using the same process. Hence, the potential benefits are found in the decrease of the variance.
- The variance reduction is obtained efficiently by averaging uncorrelated trees.
- The random choice in each partition guarantees that few variables don't dominate the regression.

Random Forests: Properties

- When each tree grows sufficiently large it is possible to decrease the bias (increasing the variance). With the average, the variance is reduced.
- The average bias is similar to the bias of each tree because each tree is built using the same process. Hence, the potential benefits are found in the decrease of the variance.
- The variance reduction is obtained efficiently by averaging uncorrelated trees.
- The random choice in each partition guarantees that few variables don't dominate the regression.

Random Forests: Performance

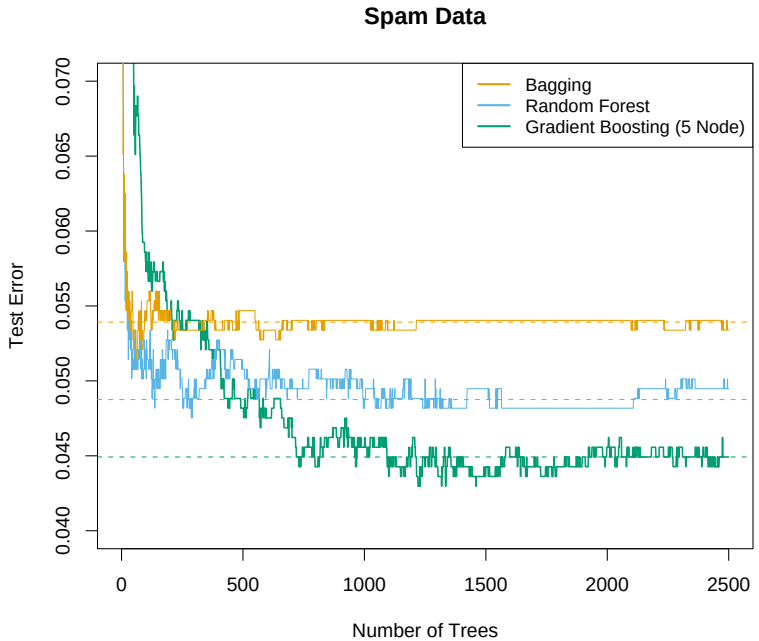


FIGURE 15.1. *Bagging, random forest, and gradient boosting, applied to the spam data. For boosting, 5-node trees were used, and the number of trees were chosen by 10-fold cross-validation (2500 trees). Each "step" in the figure corresponds to a change in a single misclassification (in a test set of 1536).*

Random Forests: Performance

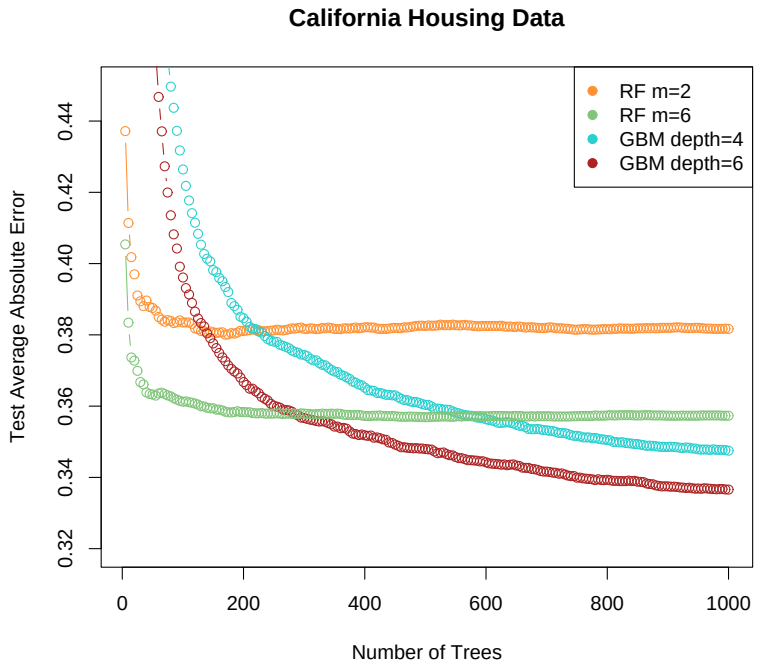


FIGURE 15.3. *Random forests compared to gradient boosting on the California housing data. The curves represent mean absolute error on the test data as a function of the number of trees in the models. Two random forests are shown, with* $m = 2$ *and* $m = 6$*. The two gradient boosted models use a shrinkage parameter* $\nu = 0.05$ *in (10.41), and have interaction depths of* 4 *and* 6*. The boosted models outperform random forests.*

Random Forests: Recommendations

- For classification $m = \sqrt{p}$ y $n_{min} = 1$.
- For regression $m = \frac{p}{3}$ y $n_{min} = 5$.
- In practice, the parameters must be calibrated. For instance, in the California example, the parameters work better with other values.

Example: Relative importance

- To measure the importance of each variable we first define the relative importance of each variable in a single tree.
- For each variable its importance can be measured as: sum the squared reduction in error at each node that the variable is used to split.

Example: Relative importance

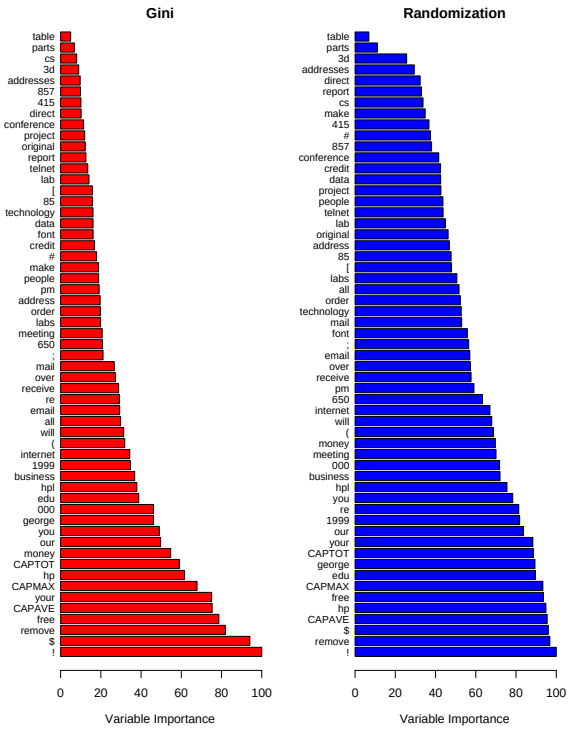


FIGURE 15.5. *Variable importance plots for a classification random forest grown on the* `spam` *data. The left plot bases the importance on the Gini splitting index, as in gradient boosting. The rankings compare well with the rankings produced by gradient boosting (Figure 10.6 on page 316). The right plot uses* OOB *randomization to compute variable importances, and tends to spread the importances more uniformly.*

Contenido

Classification Models

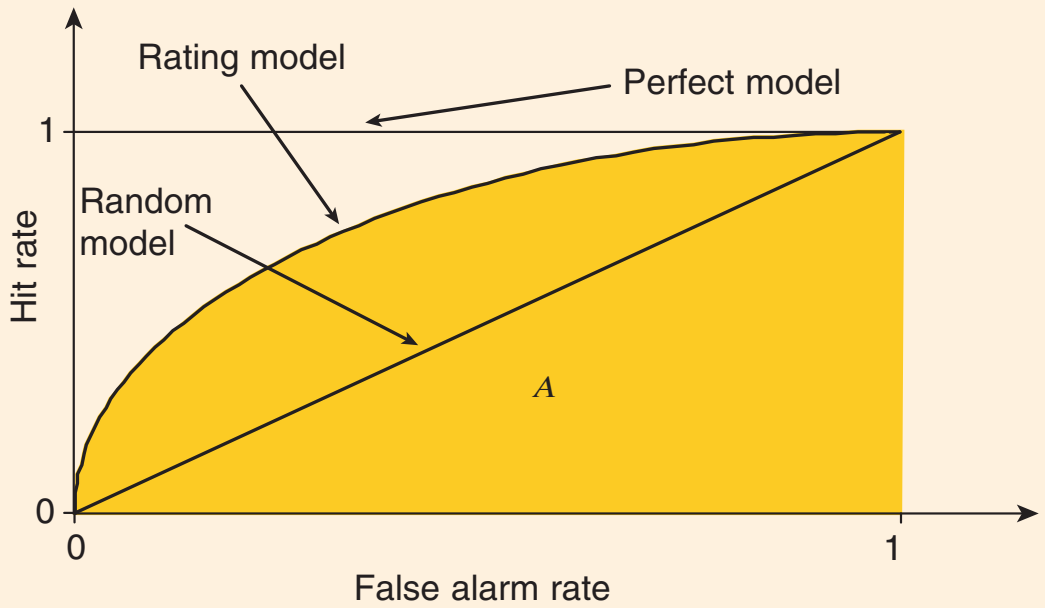
- Regression models: AIC, R^2, MAPE, etc.
- Classification models: ROC curve y calibration curve, etc.

ROC Curve

- ROC Curve and the area under the curve are important methods of validation for classification problems.

ROC Curve

3. Receiver operating characteristic curves



- Binary classification models can be extended to multi-class classification models.

ROC Curve

- Consider a good and bad scores cumulative distribution graph. The score that represents the maximum distance between these two distributions is the Kolmogorov-Smirnov distance.
- If we draw these two graphics in the same plot, we obtain the ROC curve: on the y-axis we present the bad score distribution function and on the x-axis the good score distribution function: sensitivity vs (1-specificity(x)).
- The KS distance represents the score where the horizontal distance between the ROC curve and the diagonal line is maximum (slope=1).
- Gini coefficient is two times the area between the diagonal line and the ROC Curve.
- In the ROC curve, KS is the point where the curve has a slope $= 1$ or the greater distance to the diagonal.

ROC Curve

- Consider a good and bad scores cumulative distribution graph. The score that represents the maximum distance between these two distributions is the Kolmogorov-Smirnov distance.
- If we draw these two graphics in the same plot, we obtain the ROC curve: on the y-axis we present the bad score distribution function and on the x-axis the good score distribution function: sensitivity vs (1-specificity(x)).
- The KS distance represents the score where the horizontal distance between the ROC curve and the diagonal line is maximum (slope=1).
- Gini coefficient is two times the area between the diagonal line and the ROC Curve.
- In the ROC curve, KS is the point where the curve has a slope = 1 or the greater distance to the diagonal.

ROC Curve

- Consider a good and bad scores cumulative distribution graph. The score that represents the maximum distance between these two distributions is the Kolmogorov-Smirnov distance.
- If we draw these two graphics in the same plot, we obtain the ROC curve: on the y-axis we present the bad score distribution function and on the x-axis the good score distribution function: sensitivity vs (1-specificity(x)).
- The KS distance represents the score where the horizontal distance between the ROC curve and the diagonal line is maximum (slope=1).
- Gini coefficient is two times the area between the diagonal line and the ROC Curve.
- In the ROC curve, KS is the point where the curve has a slope = 1 or the greater distance to the diagonal.

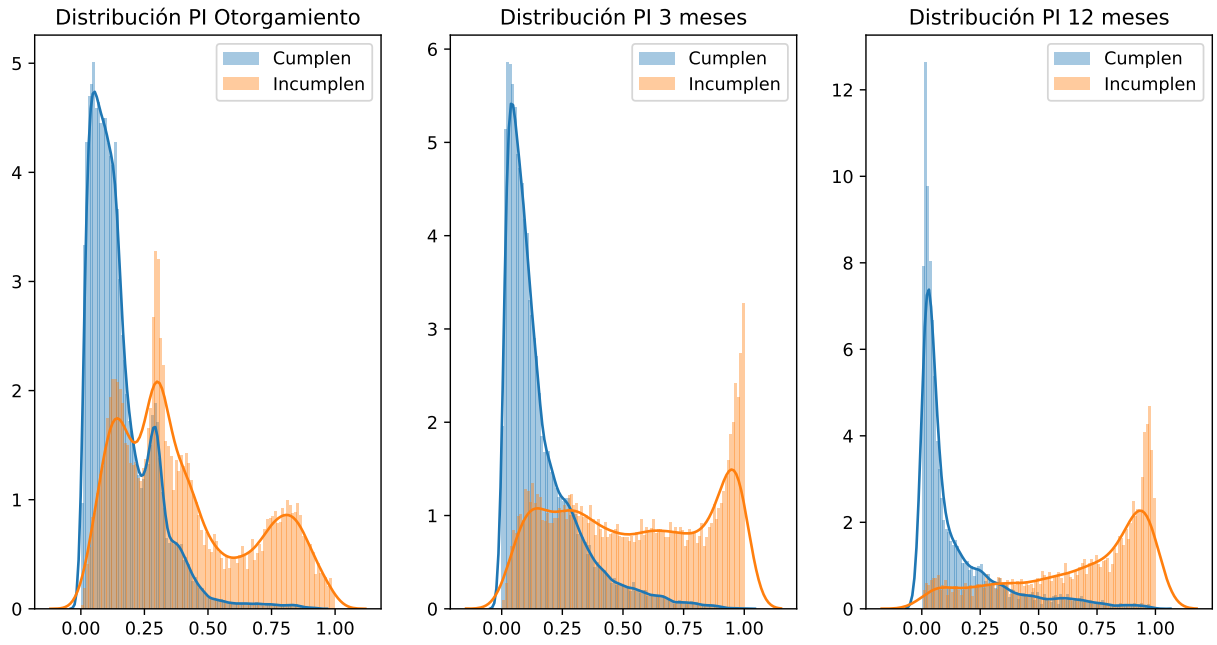
ROC Curve

- Consider a good and bad scores cumulative distribution graph. The score that represents the maximum distance between these two distributions is the Kolmogorov-Smirnov distance.
- If we draw these two graphics in the same plot, we obtain the ROC curve: on the y-axis we present the bad score distribution function and on the x-axis the good score distribution function: sensitivity vs (1-specificity(x)).
- The KS distance represents the score where the horizontal distance between the ROC curve and the diagonal line is maximum (slope=1).
- Gini coefficient is two times the area between the diagonal line and the ROC Curve.
- In the ROC curve, KS is the point where the curve has a slope = 1 or the greater distance to the diagonal.

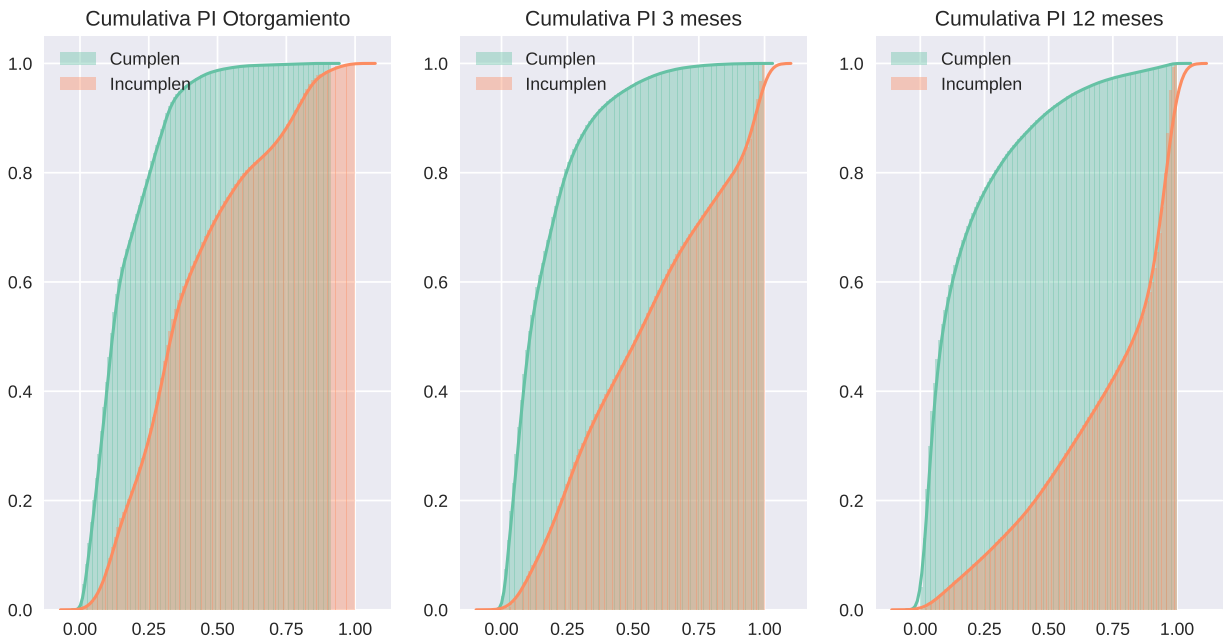
ROC Curve

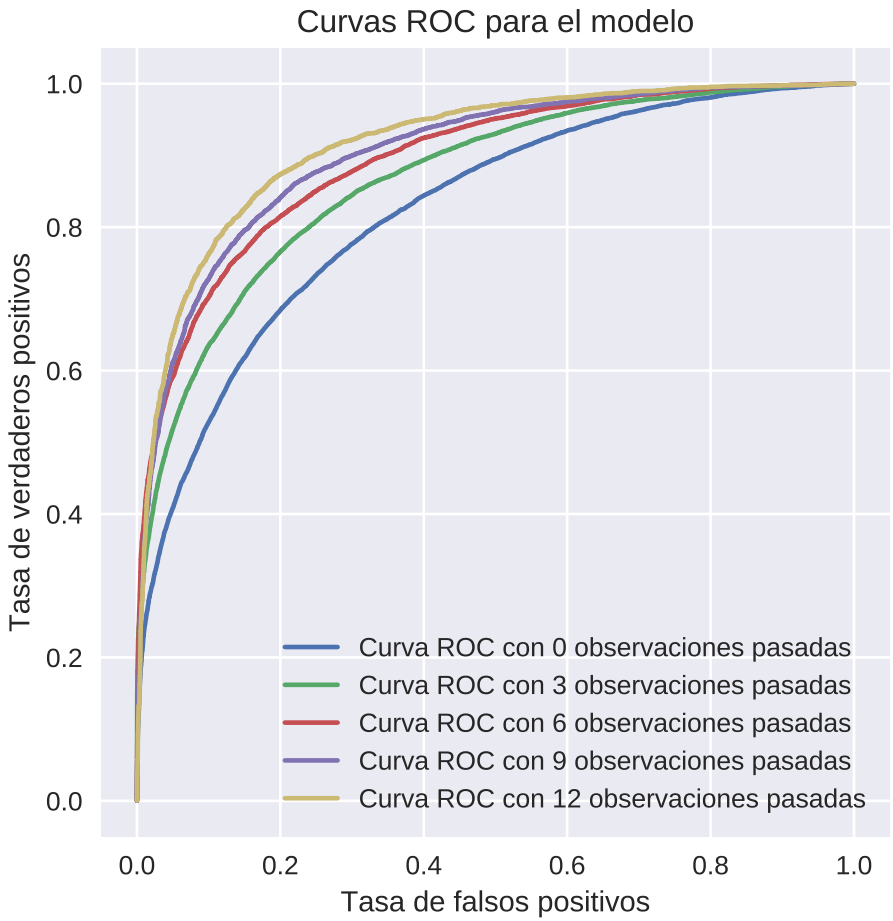
- Consider a good and bad scores cumulative distribution graph. The score that represents the maximum distance between these two distributions is the Kolmogorov-Smirnov distance.
- If we draw these two graphics in the same plot, we obtain the ROC curve: on the y-axis we present the bad score distribution function and on the x-axis the good score distribution function: sensitivity vs (1-specificity(x)).
- The KS distance represents the score where the horizontal distance between the ROC curve and the diagonal line is maximum (slope=1).
- Gini coefficient is two times the area between the diagonal line and the ROC Curve.
- In the ROC curve, KS is the point where the curve has a slope $= 1$ or the greater distance to the diagonal.

ROC Curve


Distribución PI Otorgamiento
Cumplen
Incumplen
0
1
2
3
4
5
0.00
0.25
0.50
0.75
1.00
Distribución PI 3 meses
Cumplen
Incumplen
0
1
2
3
4
5
6
0.00
0.25
0.50
0.75
1.00
Distribución PI 12 meses
Cumplen
Incumplen
0
2
4
6
8
10
12
0.00
0.25
0.50
0.75
1.00

ROC Curve


Cumulativa PI Otorgamiento
Cumplen
Incumplen
1.0
0.8
0.6
0.4
0.2
0.0
0.00
0.25
0.50
0.75
1.00
Cumulativa PI 3 meses
Cumplen
Incumplen
1.0
0.8
0.6
0.4
0.2
0.0
0.00
0.25
0.50
0.75
1.00
Cumulativa PI 12 meses
Cumplen
Incumplen
1.0
0.8
0.6
0.4
0.2
0.0
0.00
0.25
0.50
0.75
1.00


Curvas ROC para el modelo
Tasa de verdaderos positivos
Tasa de falsos positivos
0.0
0.2
0.4
0.6
0.8
1.0
Curva ROC con 0 observaciones pasadas
Curva ROC con 3 observaciones pasadas
Curva ROC con 6 observaciones pasadas
Curva ROC con 9 observaciones pasadas
Curva ROC con 12 observaciones pasadas

Calibration Curve

- Measures the error between the predicted frequencies of an event and the observed frequencies.
- In most machine learning applications, we use χ^2 test to determine the statistical significance of the error.

Calibration Curve

- Measures the error between the predicted frequencies of an event and the observed frequencies.
- In most machine learning applications, we use χ^2 test to determine the statistical significance of the error.

Calibration Curve

- Measures the error between the predicted frequencies of an event and the observed frequencies.
- In most machine learning applications, we use χ^2 test to determine the statistical significance of the error.