ETC3250/5250: Model assessment

Semester 1, 2020

Professor Di Cook

Econometrics and Business Statistics
Monash University

Week 9 (b)

```r
library(statquotes)
search_quotes(search="Holdane", fuzzy=TRUE)
```

```
## In scientific thought we adopt the simplest theory which will explain
## all the facts under consideration and enable us to predict new facts of
## the same kind. The catch in this criterion lies in the world
## ``simplest.'' It is really an aesthetic canon such as we find implicit
## in our criticisms of poetry or painting. The layman finds such a law as
## $dx/dt = K(d^2x/dy^2)$ much less simple than "it oozes," of which it is
## the mathematical statement. The physicist reverses this judgment, and
## his statement is certainly the more fruitful of the two, so far as
## prediction is concerned. It is, however, a statement about something
## very unfamiliar to the plainman, namely, the rate of change of a rate
## of change.
## --- John Burdon Sanderson Haldane (1892--1964) Possible Worlds, 1927.
```

```r
statquote(source="Box")
```

```
## A useful type of time series model is a recipe for transforming serial
## data into white noise.
## --- George E. P. Box
```

Know your data. ✈️

Quantitative or qualitative response? Predictors all quantitative? Do you have independent observations?

Know your data. ✈️

Quantitative or qualitative response? Predictors all quantitative? Do you have independent observations?

Plot your data. 🖼️

Is there a relationship between response and predictors? Is the relationship linear? Are boundaries linear?Is variability heterogeneous? Are groups distinct? Are there unusual observations?

Know your data. ✈️

Quantitative or qualitative response? Predictors all quantitative? Do you have independent observations?

Plot your data. 🖼️

Is there a relationship between response and predictors? Is the relationship linear? Are boundaries linear?Is variability heterogeneous? Are groups distinct? Are there unusual observations?

Check for missing values. 🔥

Do some variables have too many missings to use them? Do some observations have too many missings to use them? What would be a useful imputation method to fix the sporadic missing value?

Know your data. ✈️

Quantitative or qualitative response? Predictors all quantitative? Do you have independent observations?

Plot your data. 🖼️

Is there a relationship between response and predictors? Is the relationship linear? Are boundaries linear?Is variability heterogeneous? Are groups distinct? Are there unusual observations?

Fit a versatile model. 💻

Compute and plot model diagnostics. Where doesn't the model do well? How can it be refined?

Check for missing values. 🔥

Do some variables have too many missings to use them? Do some observations have too many missings to use them? What would be a useful imputation method to fix the sporadic missing value?

Know your data. ✈️

Quantitative or qualitative response? Predictors all quantitative? Do you have independent observations?

Plot your data. 🖼️

Is there a relationship between response and predictors? Is the relationship linear? Are boundaries linear?Is variability heterogeneous? Are groups distinct? Are there unusual observations?

Fit a versatile model. 💻

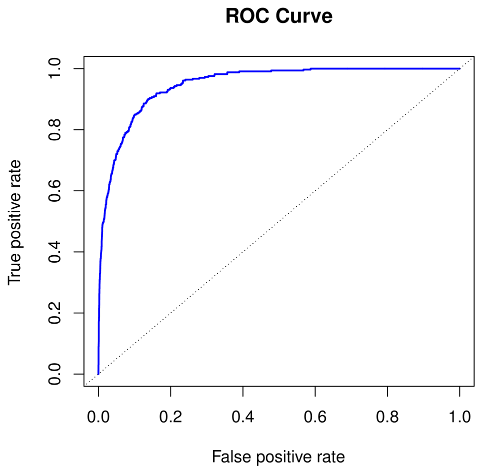
Compute and plot model diagnostics. Where doesn't the model do well? How can it be refined?

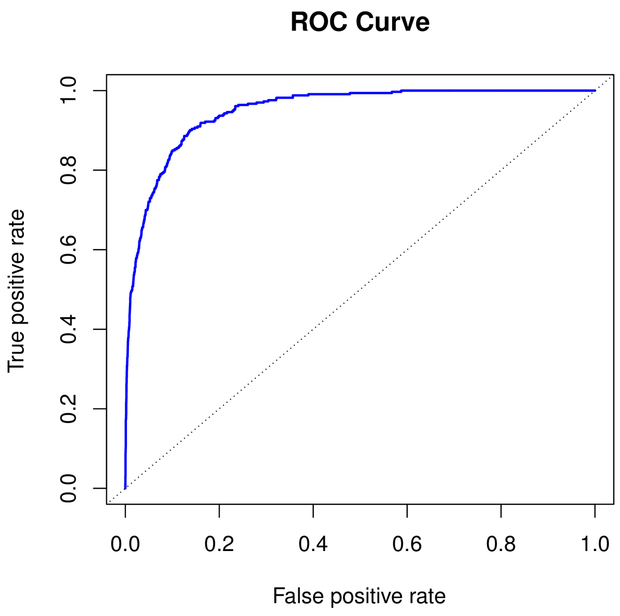
Check for missing values. 🔥

Do some variables have too many missings to use them? Do some observations have too many missings to use them? What would be a useful imputation method to fix the sporadic missing value?

ROC for classification

The ROC curve is a popular graphic for simultaneously displaying the two types of errors for all possible thresholds. It is a common method for comparing classification models. Below: ROC curve for the LDA classifier on the training set of `credit` data.





The true positive rate is the sensitivity: the fraction of defaulters that are correctly identified, using a given threshold value.

The false positive rate is 1-specificity: the fraction of non-defaulters that we classify incorrectly as defaulters, using that same threshold value.

The dotted line is "no information" classifier; class and predictor are not associated.

The ideal ROC curve hugs the top left corner, indicating a high true positive rate and a low false positive rate.

Possible outcomes with a two class classification model:

		Predicted class		
		− or Null	+ or Non-null	Total
True class	− or Null	True Neg. (TN)	False Pos. (FP)	N
	+ or Non-null	False Neg. (FN)	True Pos. (TP)	P
	Total	N^*	P^*	

+ has disease (class = 1 or "P")

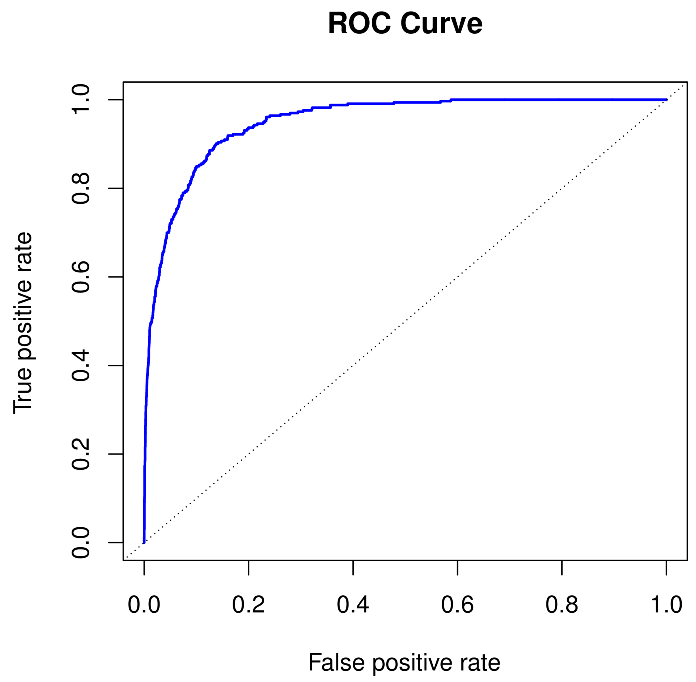
− does NOT have disease (class = 0 or "N")

True = we get it right

False = we got it wrong

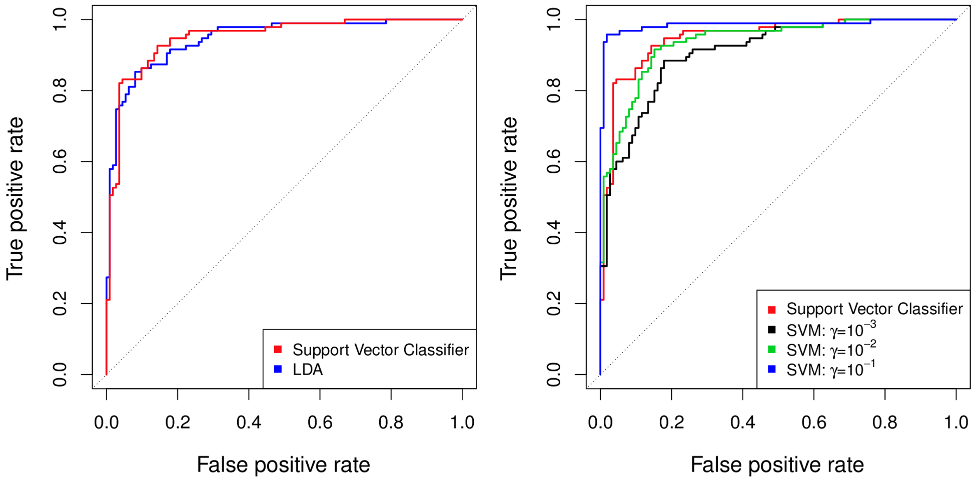
False positive rate = FP/N, also known as Type I error or 1-specificity

True positive rate = TP/P, also known as power, sensitivity, recall



If the classifier returns a prediction between 0 and 1, interpret as the probability of a positive, then threshold this at different values, e.g. 0.1, 0.2, 0.3, 0.4, 0.5, ...

ROC for classification

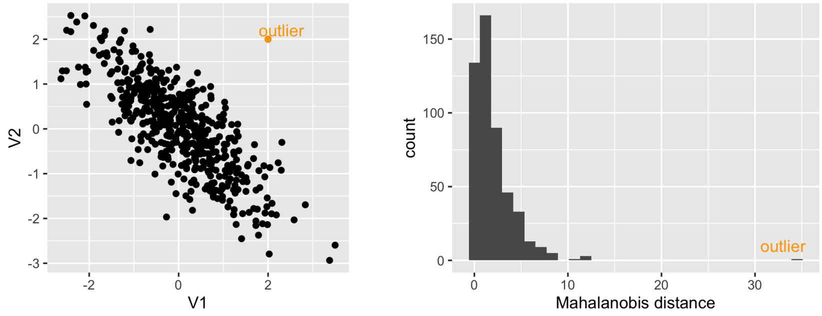


(left) LDA and SVM similar.
(right) SVM radial basis with $\gamma = 10^{-1}$ is the best.

Fig 9.10

Multivariate outliers

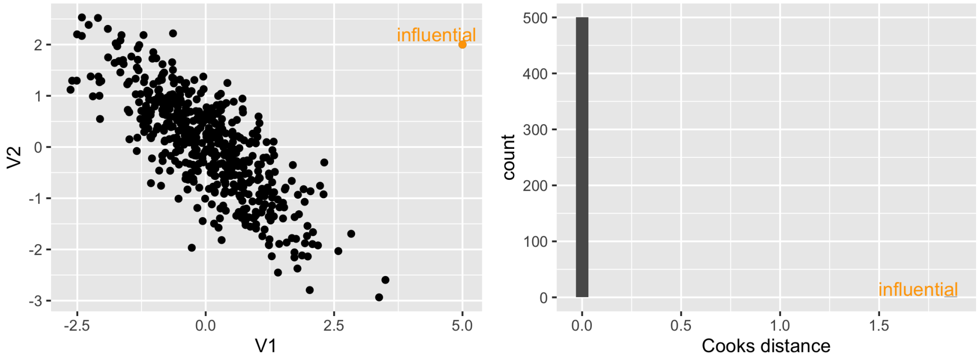
Mahalanobis distance measures the distance from the mean, relative to the variance-covariance matrix, and is useful for outlier detection:
$D^2 = (X - \mu)'\Sigma^{-1}(X - \mu)$



Related to "leverage" in regression diagnostics.

Influential observations

Cook's distance measures the change in the model estimates due to the observation: $D_i = \frac{e_i^2}{MSE \times p} \frac{h_i}{(1-h_i)^2}$ where h_i is the leverage of observation i.



Developed by Dennis Cook, University of Minnesota.

Utilising bagging

Remember the vote matrix available from random forests:

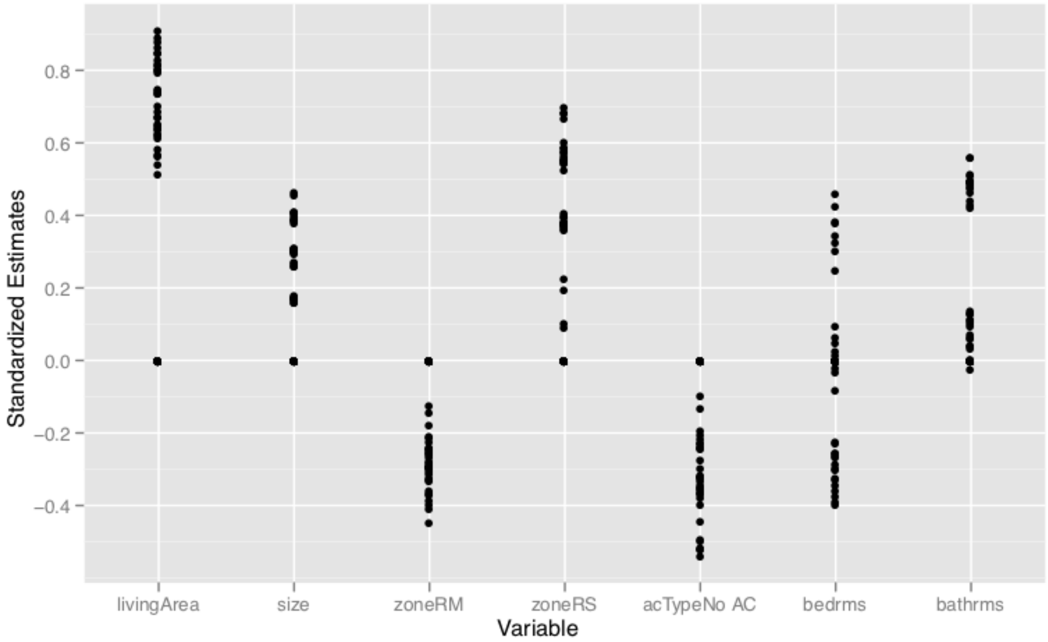
$$
\begin{aligned}
V &= (V_1 V_2 \ldots V_K) \\
&= \begin{bmatrix} p_{11} p_{12} & \ldots & p_{1K} \\ p_{21} p_{22} & \ldots & p_{2K} \\ \ldots\ldots & & \ldots \\ p_{n1} p_{n2} & \ldots & p_{nK} \end{bmatrix}
\end{aligned}
$$

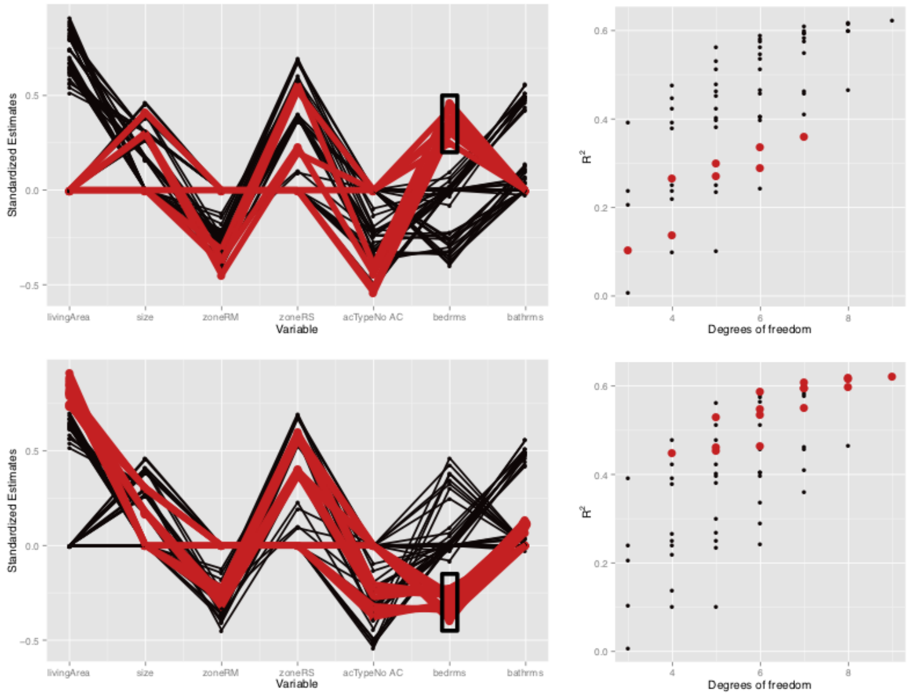
With bagging, multiple out of bag predictions produces uncertainty measure for each observation. It's possible that observations with higher uncertainty are outliers.

Variable importance

- Working with standardised variables helps, because magnitude of coefficients is then directly interpreted as importance
- Permutation approach in random forests is useful more broadly. Compare magnitude of coefficients between models built on original and permuted variable.
- Effect of one predictor with the response can depend on their relationship with one another. Called multicollinearity in regression.

All possible model fits to housing data with 7 variables, from Wickham et al (2015) Removing the Blindfold





Three typical estimates for bedrooms: big positive, close to 0, big negative.

Models with big positive coefficients for bedrooms tend to have weaker fits. They tend to occur with models that have no livingArea contribution, and more negative coefficients for zoneRM, and no air con.

Models with big negative coefficients on bedrooms tend to have stronger fits. All contrast with livingArea (high positive coefficients).

If bedrooms contribute to the model, bathrooms do not.

Model choice - robustness of conclusions

Whatever way you model the data, the interpretations should be consistent.

- Bias can explain difference in predictions between models, flexible vs inflexible can provide a spectrum on what the data predicts.
- Broad changes in a model when some cases or some variables are not used, should evoke suspicions (your "spidey sense").
- Model fit statistics are a measure of predictive power. A weak model can still be useful if there is a large cost involved.

👩‍💻 Made by a human with a computer

Slides at https://iml.numbat.space.

Code and data at https://github.com/numbats/iml.

Created using R Markdown with flair by **xaringan**, and **kunoichi** (female ninja) style.