Foundations of Statistical and Machine Learning for Actuaries
Classification, Logistic Regression and Trees

Edward (Jed) Frees, University of Wisconsin - Madison
Andres M. Villegas, University of New South Wales

July 2025

Schedule

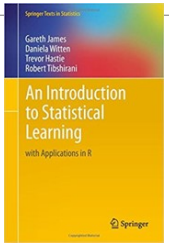
Day and Time	Presenter	Topics	Notebooks for Participant Activity
Monday Morning	Jed Jed	Welcome and Foundations; Hello to Google Colab Classical Regression Modeling	Auto Liability Claims Medical Expenditures (MEPS)
Monday Afternoon	Andrés Andrés	Resampling, cross-validation and regularisation Classification, Logistic Regression and Trees	Seattle House Sales Victoria road crash data
Tuesday	Andrés	(Tree-based) Ensembles methods and Interpretability	Victoria road crash data
Morning	Jed	Big Data, Dimension Reduction and Non-Supervised Learning	Big Data, Dimension Reduction and Non-Supervised Learning
Tuesday Afternoon	Jed Jed Fei	Neural Networks Graphic Data Neural Networks Fei Huang Thoughts on Ethics	Seattle House Prices, Claim Counts MNIST Digits Data
Wednesday Morning	Jed Jed	Recurrent Neural Networks, Text Data Artificial Intelligence, Natural Language Processing, and ChatGPT	Insurer Stock Returns
Wednesday After Lunch	Dani	Dani Bauer Insights	
Wednesday Afternoon	Andrés	Applications and Wrap-Up	

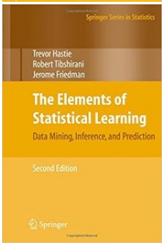
Monday Afternoon 2B - Classification, Logistic Regression and Trees

This module covers:

- Classification problems
- Model accuracy in classification problems
  - Confusion matrix
  - ROC curve
- Logistic regression
- Introduction to tree-based methods

Statistical Machine Learning: Resources





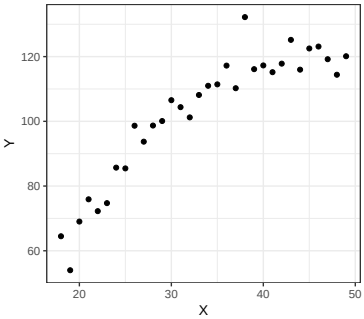
- Most of the discussion is based on this book:
  - Available at: https://www.statlearning.com/
  - Focus on intuition and practical implementation

- This book can serve as reference for those interested in the math behind the methods
- Available at: http://web.stanford.edu/~hastie/ElemStatLearn/

The discussion builds on the UNSW Course Statistical Machine Learning for Risk and Actuarial Applications (https://unsw-risk-and-actuarial-studies.github.io/ACTL3142/)

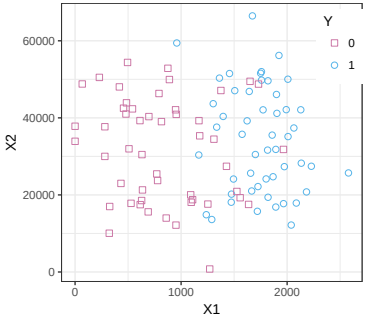
Regression vs. classification

Regression



- Y is quantitative, continuous
- Examples: Sales prediction, claim size prediction, stock price modelling

Classification



- Y is qualitative, discrete
- Examples: Fraud detection, face recognition, accident occurrence, death

Classification problems

- Coding in the binary case is simple:

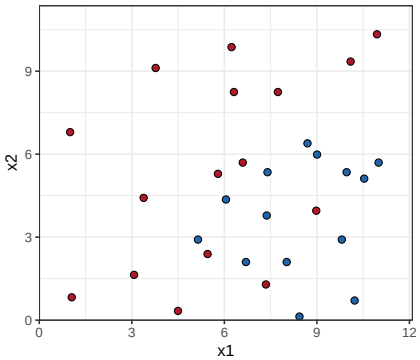
$$Y \in \{0, 1\} \Leftrightarrow Y \in \{\bullet, \bullet\}$$

- Our objective is to find a good predictive model f that can:
  1. Estimate the probability $\Pr(Y = 1|X) \in \{0, 1\}$

$$f(X) \rightarrow \bullet\bullet\bullet\bullet\bullet\bullet\bullet\bullet\bullet$$

  2. Classify observation

$$f(X) \rightarrow \hat{Y} \in \{\bullet, \bullet\}$$



Logistic regression

Extend linear regression to model binary categorical variables

$$\underbrace{\ln\left(\frac{\mathbb{P}(Y=1|X)}{1-\mathbb{P}(Y=1|X)}\right)}_{\text{log-odds}} = \underbrace{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}_{\text{linear model}}$$

Principles of Logistic Regression

- The output is binary $Y \in \{1, 0\}$
- Each case's Y variable has a probability between 0 and 1 that depends on the values of the predictors X such that

$$\mathbb{P}(Y = 1|X) + \mathbb{P}(Y = 0|X) = 1$$

- Probability can be restated as odds

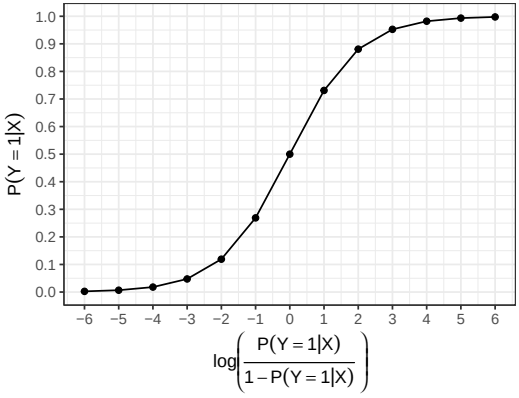
$$\text{Odds}(Y = 1|X) = \frac{\mathbb{P}(Y = 1|X)}{\mathbb{P}(Y = 0|X)} = \frac{\mathbb{P}(Y = 1|X)}{1 - \mathbb{P}(Y = 1|X)}$$

- Odds are a measure of relative probabilities

Probabilities, odds and log-odds

Goal: Transform a number between 0 and 1 into a number between $-\infty$ and $-\infty$

probability	odds	logodds
0.001	0.001	-6.907
0.250	0.333	-1.099
0.500	1.000	0.000
0.750	3.000	1.099
0.999	999.000	6.907



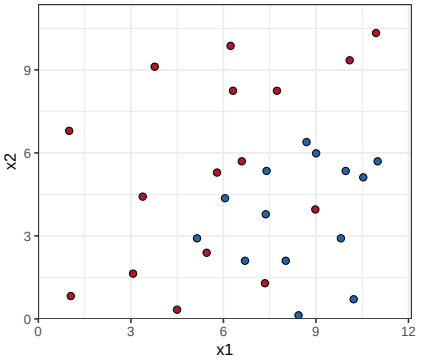
Logistic regression

- Perform regression on log-odds

$$\ln\left(\frac{\mathbb{P}(Y=1|X)}{1-\mathbb{P}(Y=1|X)}\right)=\beta_0+\beta_1X_1+\cdots+\beta_pX_p$$

- Use (training) data and maximum-likelihood estimation to produce estimates $\hat{\beta}_0,\hat{\beta}_1,\ldots\hat{\beta}_p$.
- Predict probabilities using

$$\mathbb{P}(Y=1|X)=\frac{e^{\hat{\beta}_0+\hat{\beta}_1X_1+\cdots+\hat{\beta}_pX_p}}{1+e^{\hat{\beta}_0+\hat{\beta}_1X_1+\cdots+\hat{\beta}_pX_p}}$$



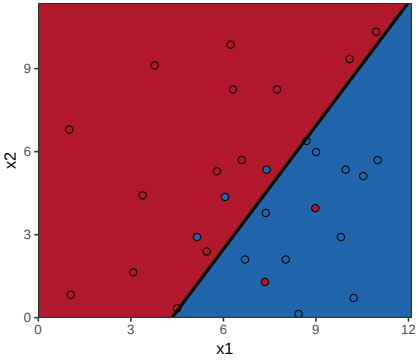
Assessing accuracy in classification problems

- We assess model accuracy using the error rate

$$\text{error rate} = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

- In our toy example with a 50% threshold

$$\text{training error rate} = \frac{5}{30} = 0.1667$$

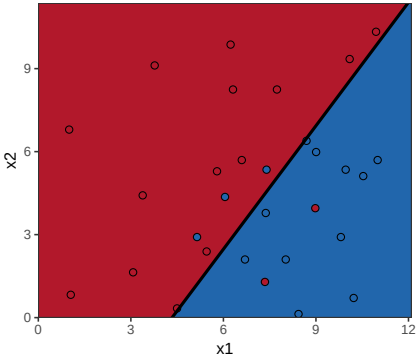


Confusion matrix (50% Threshold)

- Confusion matrix

	$Y = 0$	$Y = 1$	Total
$\hat{Y} = 0$	12	3	15
$\hat{Y} = 1$	2	13	15
Total	14	16	30

- True-Positive Rate $= \frac{13}{16} = 0.875$
- False-Positive Rate $= \frac{2}{14} = 0.1428$

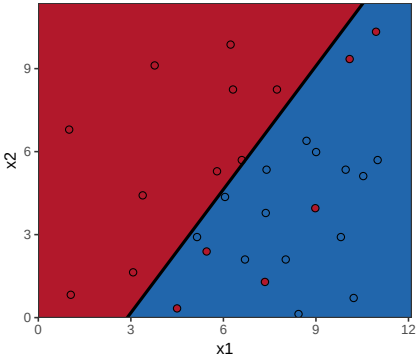


Confusion matrix (25% Threshold)

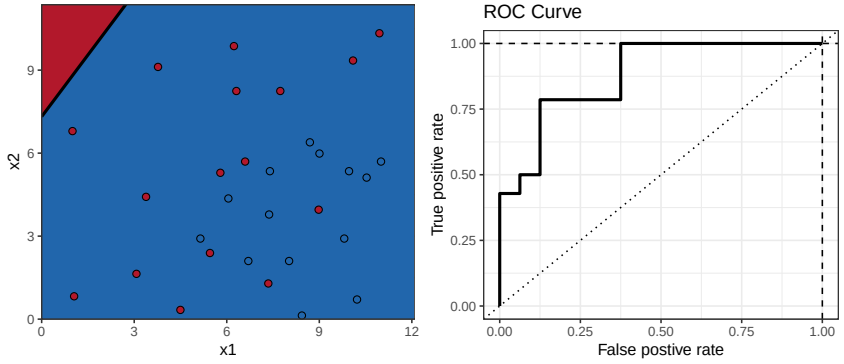
- Confusion matrix

	$Y = 0$	$Y = 1$	Total
$\hat{Y} = 0$	10	0	10
$\hat{Y} = 1$	6	16	22
Total	14	16	30

- True-Positive Rate $= \frac{16}{16} = 1$
- False-Positive Rate $= \frac{6}{14} = 0.4286$



ROC Curve and AUC



- ROC Curve: Plots the true-positive rate against the false-positive rate
- A good model will have its ROC curve hug the top-left corner more
- AUC is the area under the ROC curve: For this toy example AUC=0.8795

Can we predict if a road accident will be fatal?

Output (Y):

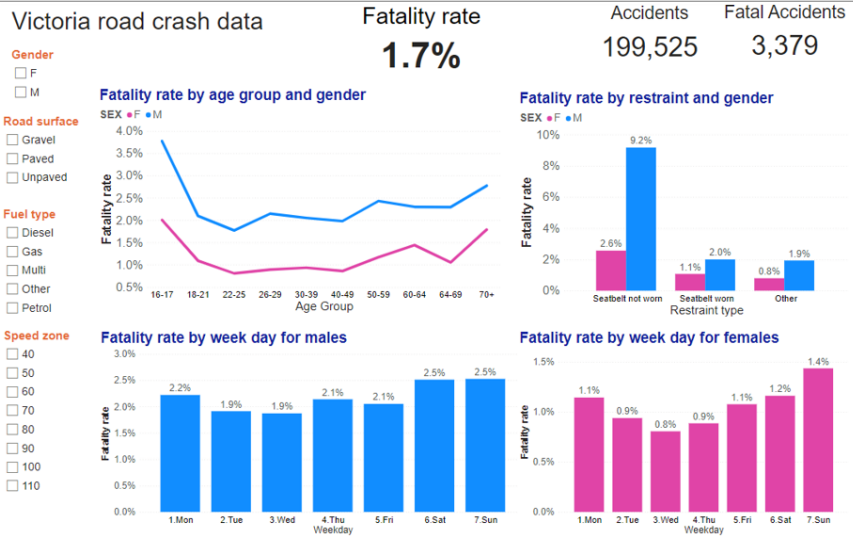
- The accident is fatal; the accident is not fatal

Input (X):

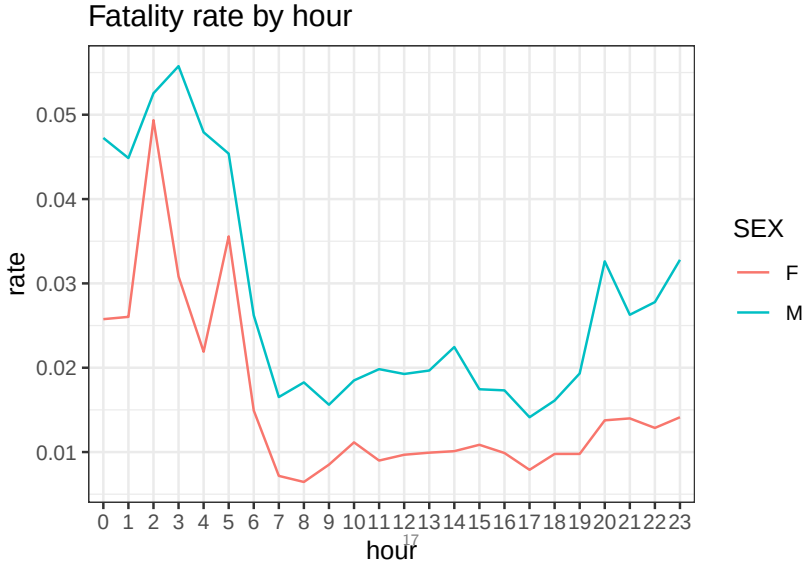
- Age of Driver
- Sex of Driver
- Time of the accident
- Weather conditions
- Type of vehicle
- ...

Source: https://discover.data.vic.gov.au/dataset/crash-stats-data-extract

VicRoads Crash Data



VicRoads Crash Data

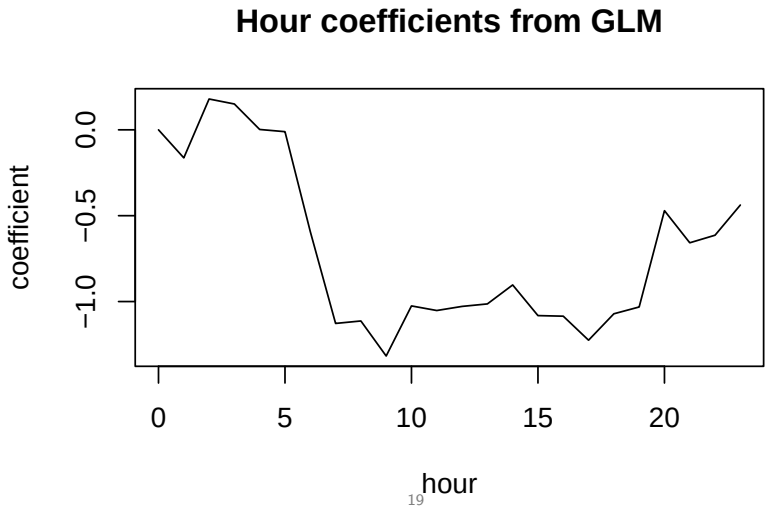


Logistic regression: VicRoads Crash Data

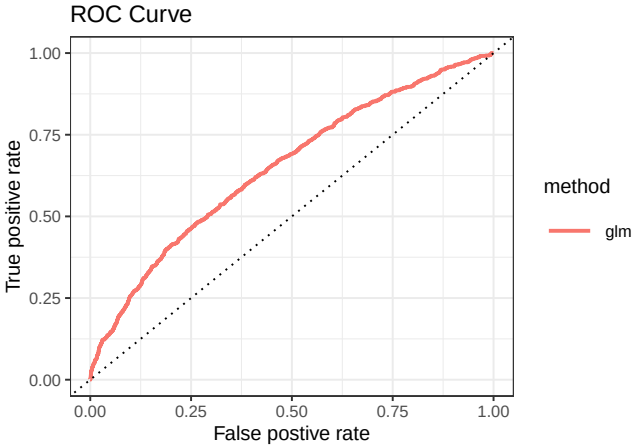
	Estimate	Pr(>|z|)
(Intercept)	-3.49	$< 2e-16^{***}$
SEXM	0.67	$< 2e-16^{***}$
SEXU	-0.38	0.59
HELMET_BELT_WORNSeatbelt not worn	1.50	$< 2e-16^{***}$
HELMET_BELT_WORNSeatbelt worn	0.12	0.01^{*}
AGE_GROUP18-21	-0.31	0.17
AGE_GROUP22-25	-0.50	0.03^{*}
⋮	⋮	⋮
AGE_GROUP70+	0.21	0.35
Weekday2.Tue	-0.13	$0.09^{\cdot}$
Weekday3.Wed	-0.20	0.01^{**}
Weekday4.Thu	-0.06	0.40
Weekday5.Fri	-0.10	0.14
Weekday6.Sat	0.00	1.00
Weekday7.Sun	0.02	0.79

Interpretation: The odds of an accident with a male driver being fatal are $\exp(0.67) = 1.95$ times higher than those of a female driver.

Logistic regression: VicRoads Crash Data



Accuracy: VicRoads Crash Data



On Test Data

- Confusion matrix

	$Y = 0$	$Y = 1$
$\hat{Y} = 0$	39324	676
$\hat{Y} = 1$	0	0

- Error Rate $= 0.01695$
- Accuracy $= 0.98305$
- AUC=0.6498

Logistic regression + regularisation

Logistic regression can be extended to include regularisation techniques to improve model performance and prevent overfitting.

- **Ridge:** Minimise on β:

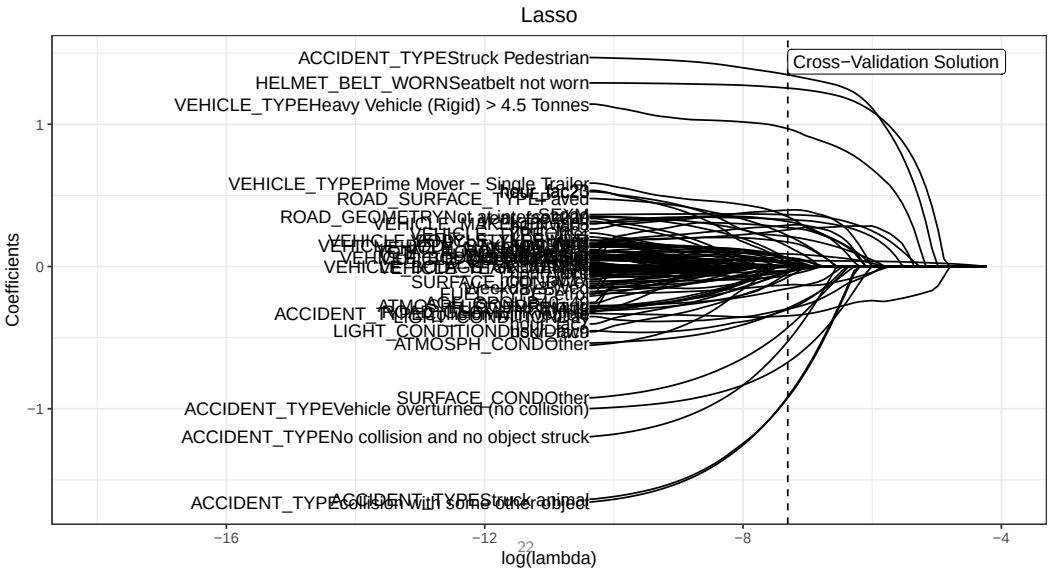
$$-\text{log-likelihood}(\beta) + \lambda \sum_{j=1}^{p} \beta_j^2$$

- **Lasso:** Minimise on β:

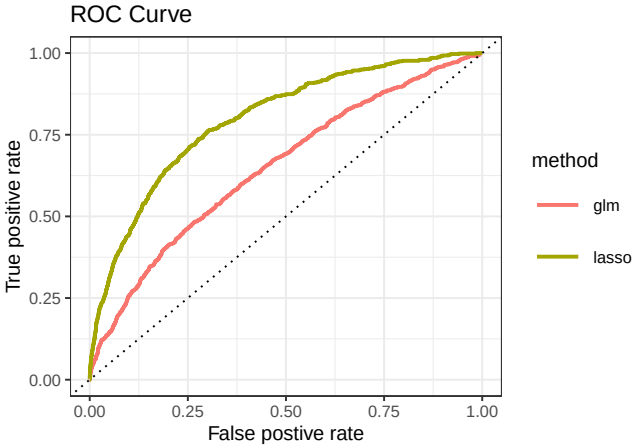
$$-\text{log-likelihood}(\beta) + \lambda \sum_{j=1}^{p} |\beta_j|$$

where λ is a tuning parameter that controls the amount of regularisation .

Logistic + Lasso: VicRoads Crash Data



Logistic + Lasso: VicRoads Crash Data (ROC)



AUC

Method	Train	Test
glm	0.663	0.650
lasso	0.809	0.797

Tree based methods

Tree based methods

Tree-based methods

- Stratify / Segment the predictor space into a number of simple regions
- The set of splitting rules can be summarised in a tree

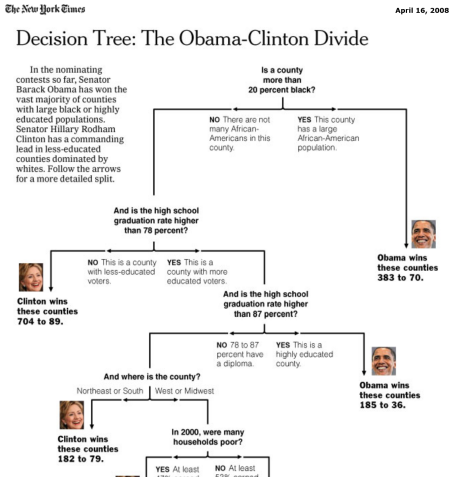
Bagging, random forests, boosting

- Ensemble methods
- Produce multiple trees
- Improve the prediction accuracy of tree-based methods
- Lose some interpretation

Tree based methods: Motivation

Trees are

- Simple
- Useful for interpretation
- Very common



Source: New York Times (2008), Decision Tree: The Obama-Clinton Divide.

Popular (IME 2023 abstracts)

Random Forests for Wildfire Insurance Applications: *Mélina Mailhot, Concordia University*

Homeowners' insurance in wildfire-prone areas can be a very risky business that some insurers may not be willing to undertake. We create an actuarial spatial model for the likelihood of wildfire occurrence over a fine grid map of North America. Several models are used, such as generalized linear models and **tree-based machine learning algorithms**. A detailed analysis and comparison of the models show a best fit using random forests. Sensitivity tests help in assessing the effect of future changes in the covariates of the model. A downscaling exercise is performed, focusing on some high-risk states and provinces. The model provides the foundation for actuaries to price, reserve, and manage the financial risk from severe wildfires.

Popular (IME 2023 abstracts)

A Machine Learning Approach to Forecasting Italian Honey Production with Tree-Based Methods: *Elia Smaniotto, University of Florence*

The Italian apiculture sector, one of the largest honey producers in Europe, has suffered considerable damage in recent years. Adverse weather conditions, occurring more frequently as climate change progresses, can be high-impact and cause the environment to be unfavourable to the bees' activity [1]. In this paper, we aim to study the effect of climatic and meteorological events on honey production. The database covers several hives, mainly located in northern Italy, and contains temperature, precipitations, geographical and meteorological measurements. We adopt **random forest and gradient boosting algorithms**, powerful and flexible **tree-based methods** to predict the honey production variation. Then, a feature importance analysis is performed to discover the main driver of honey production within the covered area. This study, which lies within the existing literature [2,3], seeks to establish the links between weather conditions and honey production, aiming to protect bees' activity better and assess potential losses for beekeepers.

Popular (IME 2023 abstracts)

Improving Business Insurance Loss Models by Leveraging InsurTech Innovation,
Emiliano Valdez, University of Connecticut

Recent transformative and disruptive developments in the insurance industry embrace various InsurTech innovations. In particular, with the rapid advances in data science and computational infrastructure, InsurTech is able to incorporate multiple emerging sources of data and reveal implications for value creation on business insurance by enhancing current insurance operations. In this paper, we unprecedentedly combine real-life proprietary insurance claims information and its InsurTech empowered risk factors describing insured businesses to create enhanced **tree-based loss models**. An empirical study in this paper shows that the supplemental data sources created by InsurTech innovation significantly help improve the underlying insurance company's internal or inhouse pricing models. The results of our work demonstrate how InsurTech proliferates firm-level value creation and how it can affect insurance product development, pricing, underwriting, claim management, and administration practice.

Popular (IME 2023 abstracts)

On the Pricing of Capped Volatility Swaps using Machine Learning Techniques, *Eva Verschueren, KU Leuven*

A capped volatility swap is a forward contract on an asset's capped, annualized realized volatility, over a predetermined period of time. The volatility swap allows investors to get a pure exposure to the volatility of the underlying asset, making the product an interesting instrument for both hedging and speculative purposes. In this presentation, we develop data-driven machine learning techniques in the context of pricing capped volatility swaps. To this purpose, we construct unique data sets comprising both the delivery price of contracts at initiation and the daily observed prices of running contracts. In order to predict future realized volatility, we explore distributional information on the underlying asset, specifically by extracting information from the forward implied volatilities and market-implied moments of the asset. The pricing performance of **tree-based machine learning models** and a Gaussian process regression model is presented in a tailored validation setting.

Popular (IME 2023 abstracts)

Integrated Design for Index Insurance, *Jinggong Zhang, Nanyang Technological University*

Weather index insurance (WII) is a promising tool for agricultural risk mitigation, but its popularity is often hindered by challenges of product design, such as basis risk, weather index selection and product complexity issues. In this paper we develop machine learning methodologies to design the statistically optimal WII to address those critical concerns in the literature and practice. The idea from **tree-based models** is exploited to simultaneously achieve weather variable selection and payout function determination, leading to effective basis risk reduction. The proposed framework is applied to an empirical study where high-dimensional weather variables are adopted to hedge soybean production losses in Iowa. Our numerical results show that the designed insurance policies are potentially viable with much lower government subsidy, and therefore can enhance social welfare.

Popular (IME 2023 abstracts)

Bayesian CART for insurance pricing, *Yaojun Zhang, University of Leeds*

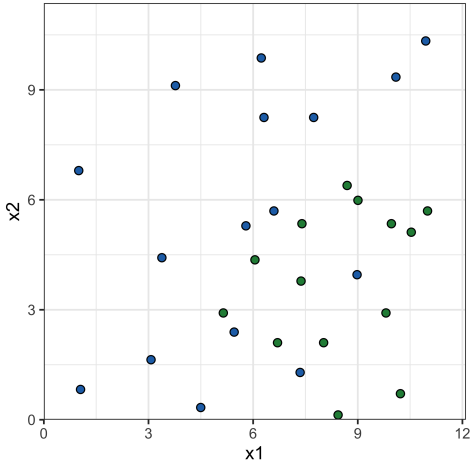
An insurance portfolio offers protection against a specified type of risk to a collection of policyholders with various risk profiles. Insurance companies use risk factors to group policyholders with similar risk profiles in tariff classes. Premiums are set to be equal for policyholders within the same tariff class which should reflect the inherent riskiness of each class. **Tree-based methods**, like the classification and regression tree (CART), have gained popularity as they can in some cases give good performance and be easily interpretable. In this talk, we discuss a Bayesian approach applied to CART models. The idea is to have the prior induce a posterior distribution that will guide the stochastic search using MCMC towards more promising trees. We shall introduce different Bayesian CART models for the insurance claims data, which include the frequency-severity model and the (zero-inflated) compound Poisson model. Some simulation and real data examples will be discussed.

Popular (IME 2023 abstracts)

Machine Learning in Long-term Mortality Forecasting, *Wenjun Zhu, Nanyang Technological University*

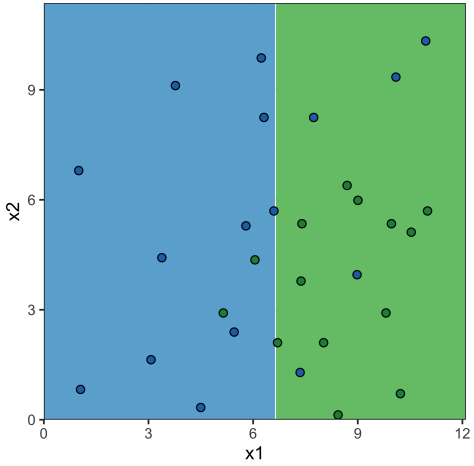
We propose a new machine learning-based framework for long-term mortality forecasting. Based on ideas of neighbouring prediction, **model ensembling, and tree boosting**, this framework can significantly improve the prediction accuracy of long-term mortality. In addition, the proposed framework addresses the challenge of a shrinking pattern in long-term forecasting with information from neighbouring ages and cohorts. An extensive empirical analysis is conducted using various countries and regions in the Human Mortality Database. Results show that this framework reduces the mean absolute percentage error (MAPE) of the 20-year forecasting by almost 50% compared to classic stochastic mortality models, and it also outperforms deep learning-based benchmarks. Moreover, including mortality data from multiple populations can further enhance the long-term prediction performance of this framework.

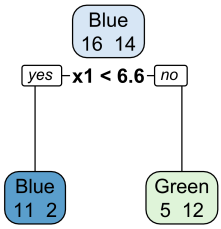
Growing a Tree I



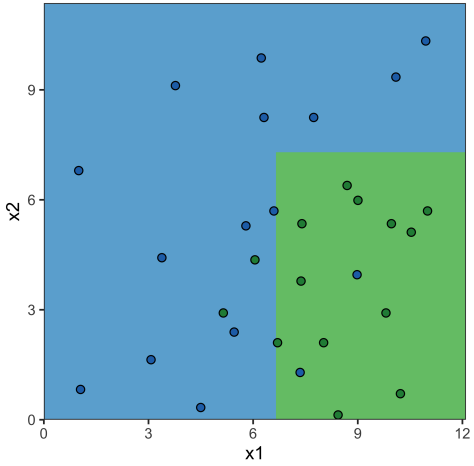
Blue
16 14

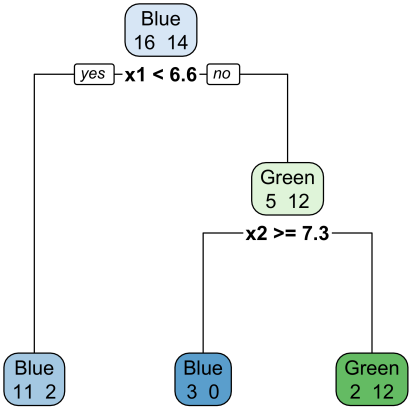
Growing a Tree II



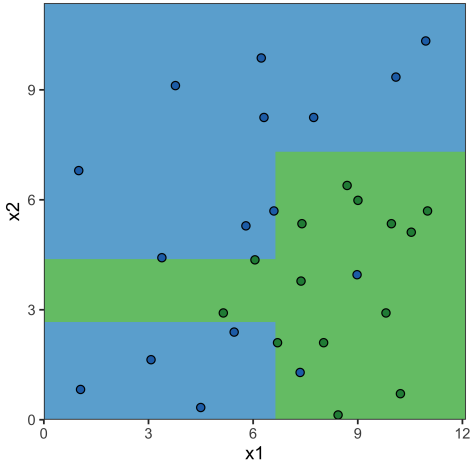


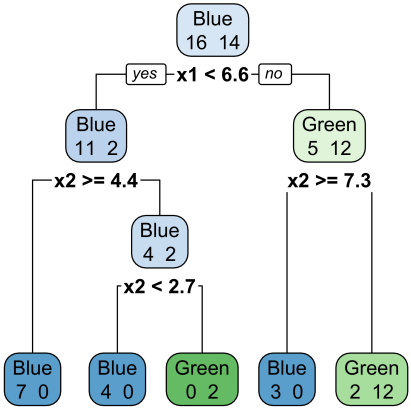
Growing a Tree III





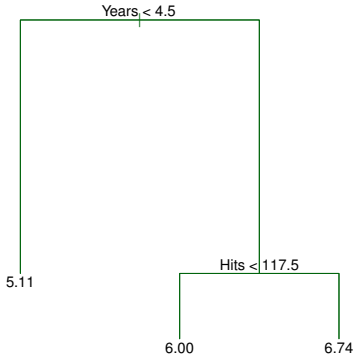
Growing a Tree IV





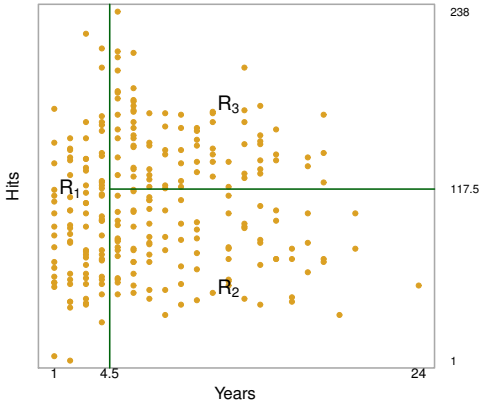
Regression Trees - Baseball Example

`Hitters` data set: predict a baseball player's log `Salary` based on `Years` and `Hits`



Source: ISLR2 Figure 8.1.

Regression Trees - Baseball Example



Source: ISLR2 Figure 8.2.

Tree regions & predictions

A decision tree is made by:

1. Dividing the predictor space (i.e. the set of possible values for $X_1, X_2, \ldots, X_p$) into J distinct and non-overlapping regions, $R_1, R_2, \ldots, R_J$,
2. Making the same prediction for every observation that falls into the region R_j
   - the mean response for the training data in R_j (regression trees)
   - the mode response for the training data in R_j (classification trees)

Region	Predicted salaries
$R_1 = \{X \vert \texttt{Years} < 4.5\}$	$\$1,000 \times \mathrm{e}^{5.107} = \$165,174$
$R_2 = \{X \vert \texttt{Years} \geq 4.5, \texttt{Hits} < 117.5\}$	$\$1,000 \times \mathrm{e}^{5.999} = \$402,834$
$R_3 = \{X \vert \texttt{Years} \geq 4.5, \texttt{Hits} \geq 117.5\}$	$\$1,000 \times \mathrm{e}^{6.740} = \$845,346$

Recursive Binary Splitting

Construct Regions

- Divide the predictor space into high-dimensional rectangles, or boxes
- The goal is to find boxes $R_1, R_2, \cdots, R_J$ that minimise

$$\text{RSS} = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

  where $\hat{y}_{R_j}$ is the mean response for the training observations within the jth box
- Computationally unfeasible to consider every possible partition
  - take a top-down, greedy approach...

Recursive Binary Splitting I

- Consider a splitting variable j and split point s

$$R_1(j,s) = \{X|X_j \leq s\} \quad \text{and} \quad R_2(j,s) = \{X|X_j > s\}$$

- Find the splitting variable j and split point s that solve

$$\min_{j,\ s}\left[\min_{c_1}\sum_{x_i\in R_1(j,\ s)}(y_i-c_1)^2+\min_{c_2}\sum_{x_i\in R_2(j,\ s)}(y_i-c_2)^2\right]$$

where the inner mins are solved by

$$\hat{c}_1 = \text{ave}(y_i|x_i \in R_1(j,s)) \quad \text{and} \quad \hat{c}_2 = \text{ave}(y_i|x_i \in R_2(j,s))$$

Recursive Binary Splitting II

- Scan through all of the inputs
  - for each splitting variable, the split point s can be determined very quickly
  - The overall solution for this branch (i.e. selection of j) follows.
- Partition the data into the two resulting regions
- Repeat the splitting process on each of the two regions
- Continue the process until a stopping criterion is reached

Classification Trees

Very similar to a regression tree, except:

- Predict that each observation belongs to the most commonly occurring class of training observations in the region to which it belongs
- RSS cannot be used as a criterion for making the binary splits, instead use a measure of node purity:

Gini index, or

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

cross-entropy

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \ln(\hat{p}_{mk})$$

where

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)\,.$$

Discussion Questions

- Why binary splitting?
- What about splitting of categorical predictors?

How large should we grow the tree?

- What's the problem if the tree is too large?
- What if the tree is too small?

Pruning a Tree

Pruning

- Wish to have a tree that has a good balance of variance vs bias.
- A decision rule of considering the decrease in RSS at each step/split (vs a threshold) is too short sighted.
- Alternate approach of growing a large tree then pruning back to obtain a subtree is a better strategy.
- Cross validation of each possible subtree is however very cumbersome.
- An alternative approach is cost complexity pruning (also known as weakest link pruning)

Cost-Complexity Pruning

Define a subtree $T \subset T_0$ to be any tree than can be obtained by pruning T_0 (a fully-grown tree)

- Terminal node m represents region R_m
- $|T|$: number of terminal nodes in T

Define the cost complexity criterion

$$\text{Total cost} = \text{Measure of Fit} + \text{Measure of Complexity}$$

$$C_\alpha(T) = \sum_{m=1}^{|T|} \sum_{i \in R_m} (y_i - \hat{y}_m)^2 + \alpha |T|$$

where $\hat{y}_m$ is the mean y_i in the mth leaf and α controls the tradeoff between tree size and goodness of fit.

Cost-Complexity Pruning

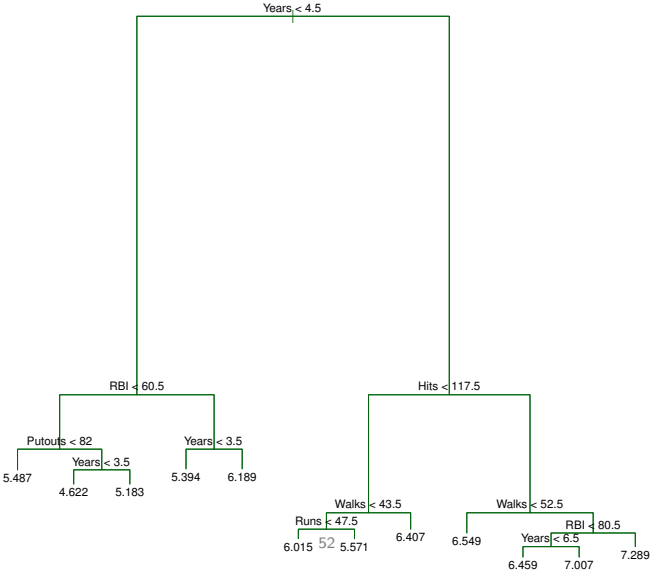
For each α, we want to find the subtree $T_\alpha \subseteq T_0$ that minimises $C_\alpha(T)$

- How to find T_α?
  - "weakest link pruning"
    - For a particular α, find the subtree T_α such that the cost complexity criterion is minimised
- How to choose α?
  - cross-validation

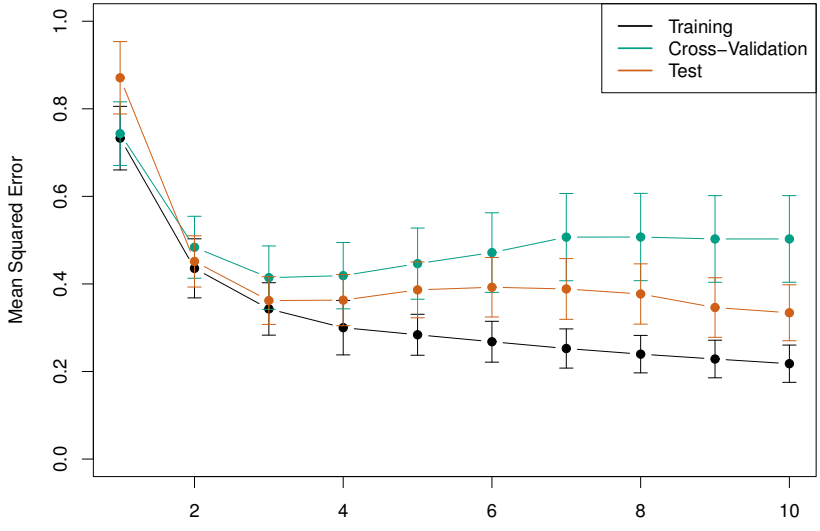
Tree Algorithm Summary

1. Use recursive binary splitting to grow a large tree on the training data
   - stop only when each terminal node has fewer than some minimum number of observations
2. Apply cost complexity pruning to the large tree to obtain a sequence of best subtrees, as a function of α
   - there is a unique smallest subtree T_α that minimises $C_\alpha(T)$
3. Use K-fold cross-validation to choose α
4. Return the subtree from Step 2 that corresponds to the chosen value of α

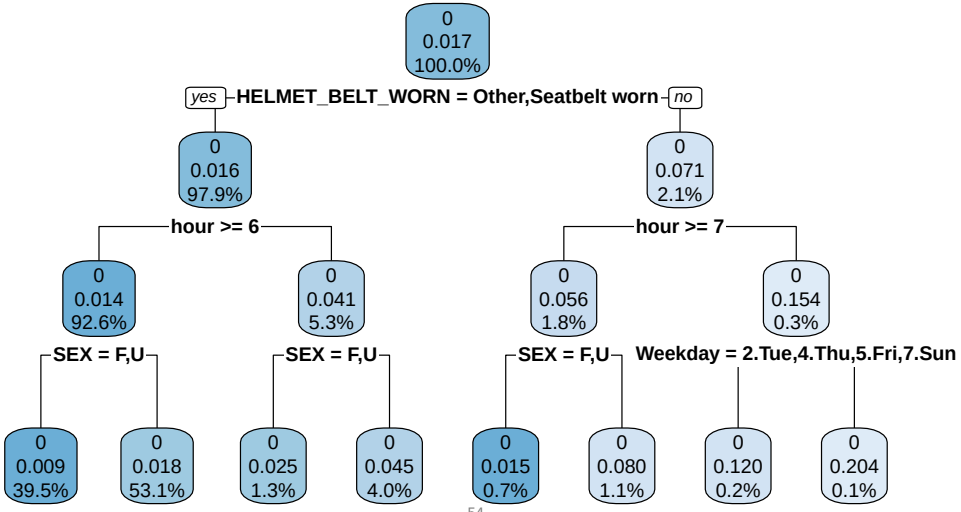
Unpruned `Hitters` tree



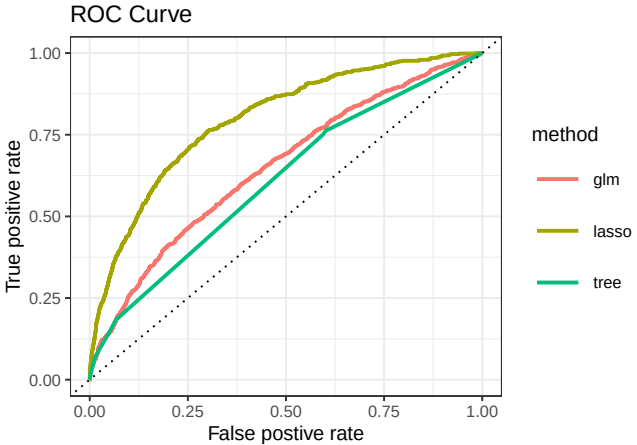
CV to pick α (equiv., $|T|$)



Tree: VicRoads Crash Data



Tree: VicRoads Crash Data (ROC)



AUC

Method	Train	Test
glm	0.663	0.650
lasso	0.809	0.797
tree	0.629	0.610

Advantages and disadvantages of Trees

Advantages

- Easy to explain
- (Mirror human decision making)
- Graphical display
- Easily handle qualitative predictors

Disadvantages

- Low predictive accuracy compared to other regression and classification approaches
- Can be very non-robust

Is there a way to improve the predictive performance of trees?

- Ensemble methods
- Bagging, random forest, boosting

Session 2B - Classification, Logistic Regression and Trees - Summary

We have discussed key concepts in classification problems

- Logistic Regression
- Assessing model accuracy
  - Confusion matrix
  - ROC curve, AOC
- Tree-based methods
  - Classification and Regression trees
  - Recursive binary splitting
  - Cost-complexity pruning
- During lab, participant may follow the notebook VicRoads Crash