Course wrap-up

STAT 471

December 9, 2021

Where we are

- ✓ **Unit 1:** Intro to modern data mining
- ✓ **Unit 2:** Tuning predictive models
- ✓ **Unit 3:** Regression-based methods
- ✓ **Unit 4:** Tree-based methods
- ✓ **Unit 5:** Deep learning

Where we are

✓ **Unit 1:** Intro to modern data mining

✓ **Unit 2:** Tuning predictive models

✓ **Unit 3:** Regression-based methods

✓ **Unit 4:** Tree-based methods

✓ **Unit 5:** Deep learning

Today's lecture:

- Final project suggestions
- Looking back at STAT 471
- Looking beyond STAT 471

Final project suggestions

Final project suggestions

- How to formulate my question?

Final project suggestions

- How to formulate my question?
  - Defining your project in the language of STAT 471

Final project suggestions

- How to formulate my question?
  - Defining your project in the language of STAT 471
  - Avoiding pitfalls: Dependent observations, class imbalance

Final project suggestions

- How to formulate my question?
  - Defining your project in the language of STAT 471
  - Avoiding pitfalls: Dependent observations, class imbalance
- How to choose the features to include in the model?

Final project suggestions

- How to formulate my question?
  - Defining your project in the language of STAT 471
  - Avoiding pitfalls: Dependent observations, class imbalance
- How to choose the features to include in the model?
  - Distinguishing features from responses in disguise

Final project suggestions

- How to formulate my question?
  - Defining your project in the language of STAT 471
  - Avoiding pitfalls: Dependent observations, class imbalance
- How to choose the features to include in the model?
  - Distinguishing features from responses in disguise
  - Hand-picking versus letting the model decide

Final project suggestions

- How to formulate my question?
  - Defining your project in the language of STAT 471
  - Avoiding pitfalls: Dependent observations, class imbalance
- How to choose the features to include in the model?
  - Distinguishing features from responses in disguise
  - Hand-picking versus letting the model decide
- What about feature correlation?

Final project suggestions

- How to formulate my question?
  - Defining your project in the language of STAT 471
  - Avoiding pitfalls: Dependent observations, class imbalance
- How to choose the features to include in the model?
  - Distinguishing features from responses in disguise
  - Hand-picking versus letting the model decide
- What about feature correlation?
  - Consequences of feature correlation

Final project suggestions

- How to formulate my question?
  - Defining your project in the language of STAT 471
  - Avoiding pitfalls: Dependent observations, class imbalance
- How to choose the features to include in the model?
  - Distinguishing features from responses in disguise
  - Hand-picking versus letting the model decide
- What about feature correlation?
  - Consequences of feature correlation
  - Visualizing feature correlation using `ggcorrplot`

Looking back at STAT 471

Themes of the class, and a lingering question

1. Regression and classification
2. Training predictive models
3. Model complexity
4. Bias-variance trade-off
5. Model selection and model assessment
6. Interpretability of predictive models

Conceptual

7. R programming and software tools
8. Working with data

Practical

Themes of the class, and a lingering question

1. Regression and classification
2. Training predictive models
3. Model complexity
4. Bias-variance trade-off
5. Model selection and model assessment
6. Interpretability of predictive models

Conceptual

7. R programming and software tools
8. Working with data

Practical

Lingering question: What is the best prediction method?

Theme: Regression and classification

Prediction methods vary based on the response type:

- Regression: continuous responses
- Classification: discrete responses (binary or multi-class)

Most methods have versions for regression and classification, e.g. linear regression and logistic regression.

Classification methods are indirect in the sense that they predict probabilities of each class. They are also a little more fussy; need to make sure probabilities are between 0 and 1, class imbalance, misclassification error versus Gini index,…

Many of the same intuitions apply for regression and classification.

Theme: Training predictive models

Define class of predictive models $f_\beta(X)$ indexed by some parameter vector β.

Find member of this class that best fits the training data, as measured by the loss function L of predictions given true responses, possibly regularized:

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f_\beta(X_i)) + \lambda \cdot \text{penalty}(\beta).$$

Theme: Training predictive models

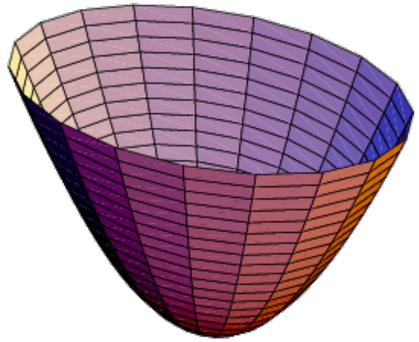
Define class of predictive models $f_\beta(X)$ indexed by some parameter vector β.

Find member of this class that best fits the training data, as measured by the loss function L of predictions given true responses, possibly regularized:

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f_\beta(X_i)) + \lambda \cdot \text{penalty}(\beta).$$

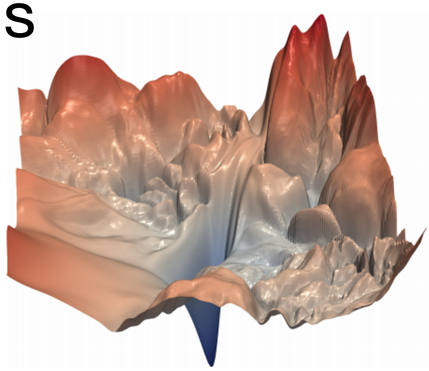
Convex (optimization is easy)

- Linear and logistic regression
- Linear and logistic regression with ridge or lasso penalties



Not convex (optimization is hard)

- Tree-based methods
- Neural networks



https://arxiv.org/abs/1712.09913

Theme: Model complexity

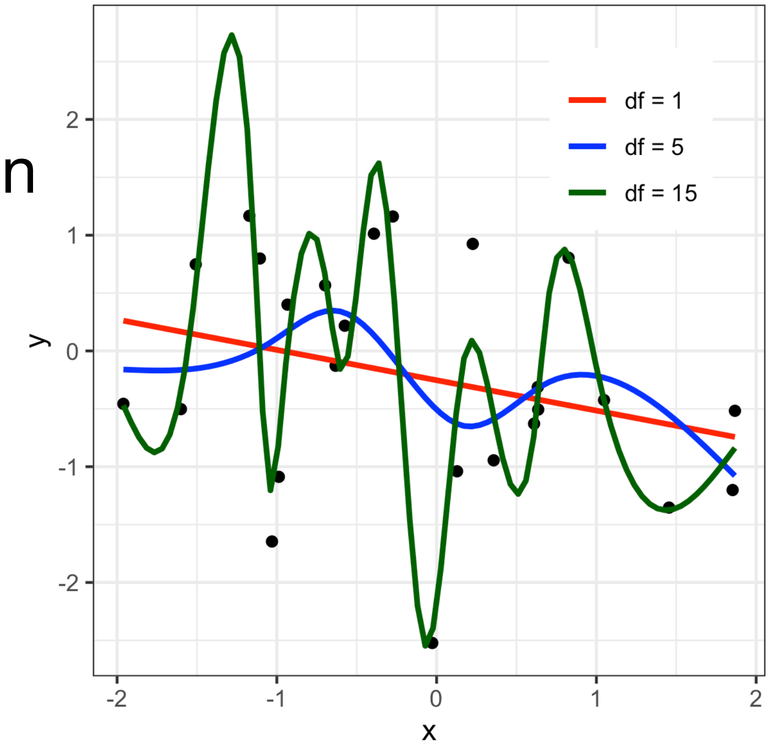
One of the key characteristics of a predictive model is its complexity: how flexibly does it fit the training data?

How is model complexity defined? Depends on model:

- Number of parameters in linear or logistic regression
- Depth of a decision tree
- Number of neighbors used in K-nearest-neighbors

How is model complexity controlled (regularization)?

- Explicit regularization via penalization (lasso, ridge)
- Implicit regularization, e.g. sub-sampling features during random forest model training



Theme: Bias-variance trade-off

Consider sampling many different training sets.

- Bias: How far off are predictions on average?
- Variance: How much do the predictions wobble around?

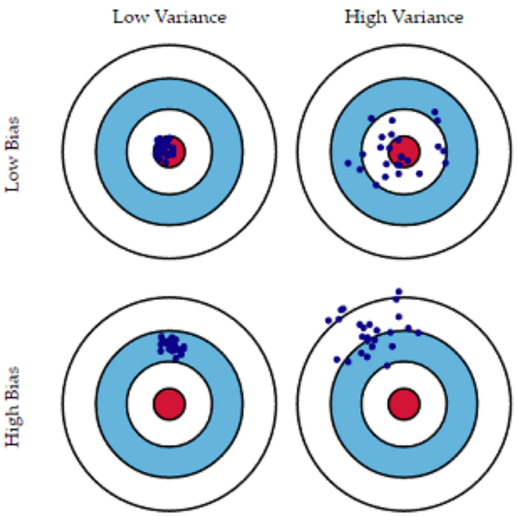
Higher model complexity leads to less bias but more variance.

$$\text{Prediction error} = \text{Bias}^2 + \text{Variance}.$$

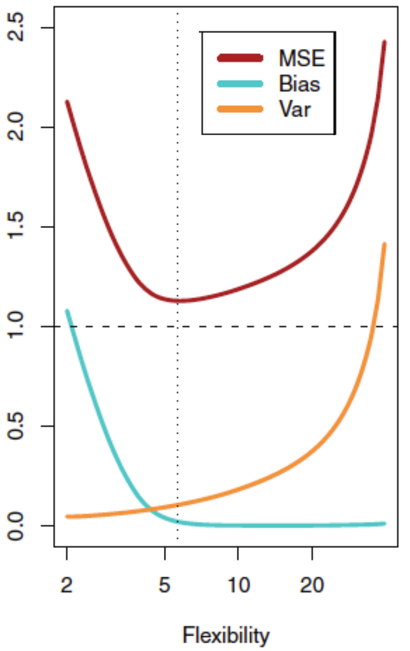
Overfitting: complex models' sensitivity to noise in the training data (high variance) → low training error but high test error.

Variance increases as noise variance increases, model flexibility increases, training sample size decreases.

Lower noise or larger sample size means you can afford more complex model (think of deep learning).



https://www.listendata.com/2017/02/bias-variance-tradeoff.html



Theme: Model selection and model assessment

Theme: Model selection and model assessment

Model assessment

- Predictive models are assessed based on test data that is separate from the data used to train these models.
- Different criteria are used to quantify the accuracy of predictions, like RMSE, misclassification error, AUC, and confusion matrix.
- Model assessment can be subtle for classification problems.

Theme: Model selection and model assessment

Model assessment

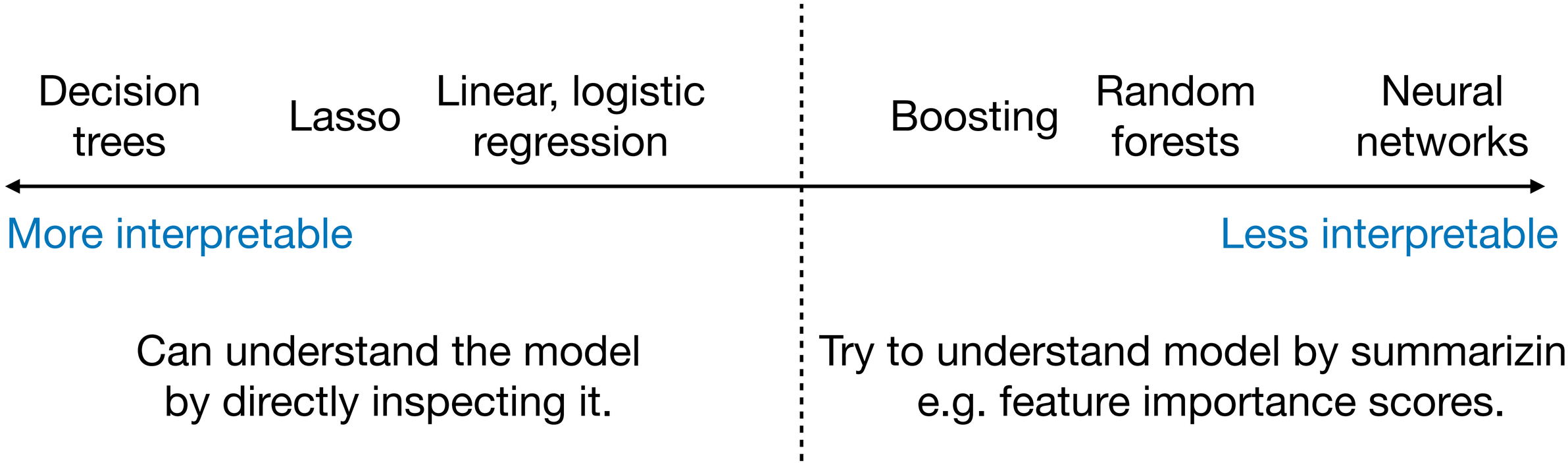
- Predictive models are assessed based on test data that is separate from the data used to train these models.
- Different criteria are used to quantify the accuracy of predictions, like RMSE, misclassification error, AUC, and confusion matrix.
- Model assessment can be subtle for classification problems.

Model selection

- Main tool: Cross-validation, which mimics the train-test split using folds.
- Other schemes for model assessment: validation set approach and out-of-bag error, the latter for random forests.
- One-standard-error rule reflects preference for simpler models.

Theme: Interpretability of predictive models

We want to understand how our predictive model is arriving at its conclusions.



Theme: R programming and software tools

Theme: R programming and software tools

Tools:

- tidyverse is a really clean way to import, clean, transform, and visualize data.
- R is well suited for data science; lots of packages available to analyze data.
- R Markdown is a nice way to integrate text, code, and output.
- Git and GitHub facilitate version control, collaboration, and sharing.
- Docker allows pre-packaged, self-contained computing environments.

Theme: R programming and software tools

Tools:

- tidyverse is a really clean way to import, clean, transform, and visualize data.
- R is well suited for data science; lots of packages available to analyze data.
- R Markdown is a nice way to integrate text, code, and output.
- Git and GitHub facilitate version control, collaboration, and sharing.
- Docker allows pre-packaged, self-contained computing environments.

Lessons learned:

- Programming takes patience, attention to detail, and lots of Googling.
- Each R package and each software tool has its own quirks and limitations.
- With practice, these programming and software tools can be very powerful.

Theme: Working with data

- Though increasingly abundant, data are still a precious resource, more of which gives better predictions.
- Especially in the real world, data are messy and require cleaning.
- Exploratory data analysis and visualization can reveal a lot about a dataset.
- The most successful analyses couple statistical intuition and data intuition.
- Ultimate goal of data science is to create knowledge and/or make decisions; we must make conclusions relevant to the underlying real-world problem.

A lingering question: What is the best method?

A lingering question: What is the best method?

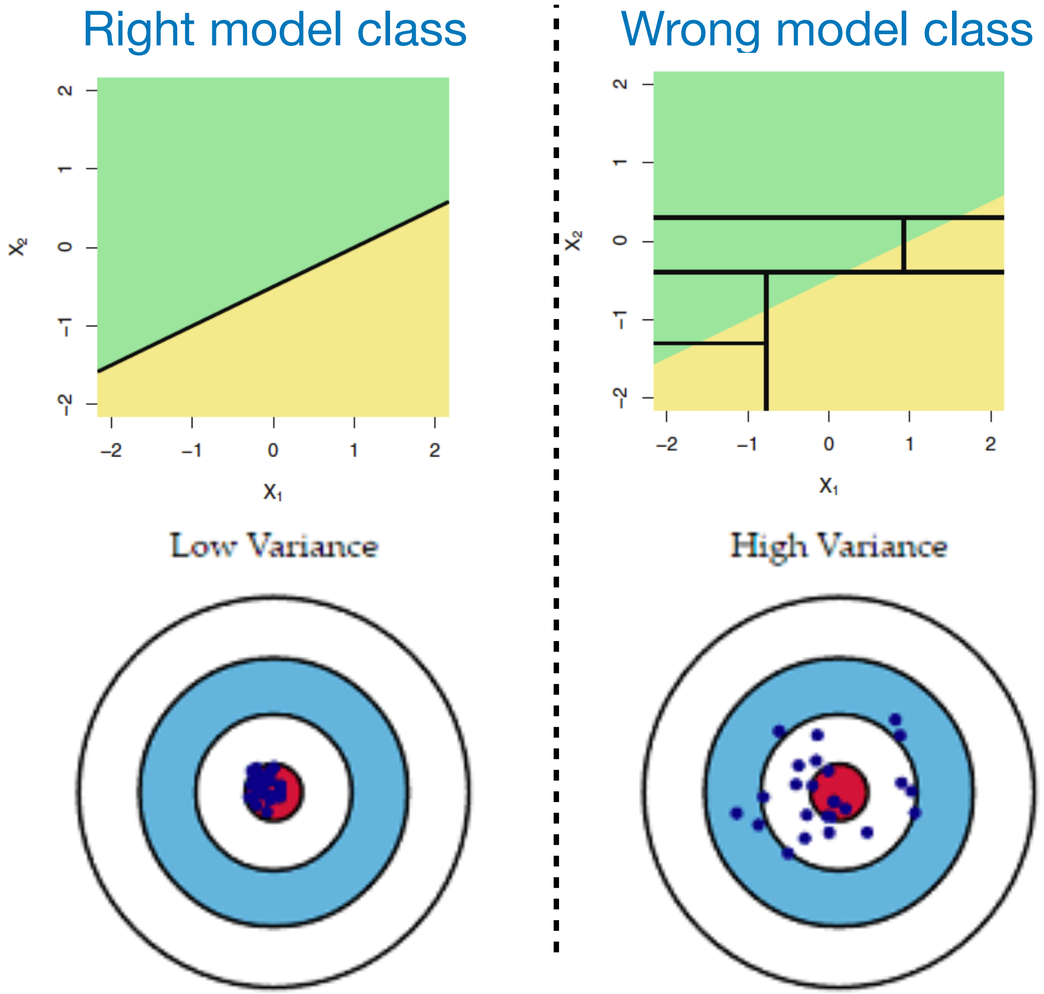
There is no best method. Different methods will work better in different situations.

A lingering question: What is the best method?

There is no best method. Different methods will work better in different situations.

Each prediction method “has in mind” an underlying model class, e.g. linear models for linear regression versus piece-wise constant models for trees.

If true feature-response relationship matches model class our method “has in mind,” will take fewer parameters (less variance) to fit underlying trend (less bias).



https://www.listendata.com/2017/02/bias-variance-tradeoff.html

A lingering question: What is the best method?

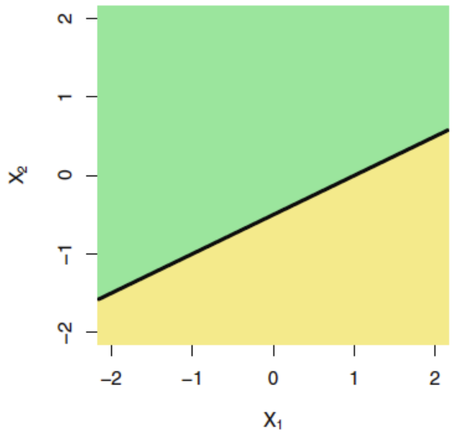
There is no best method. Different methods will work better in different situations.

Each prediction method “has in mind” an underlying model class, e.g. linear models for linear regression versus piece-wise constant models for trees.

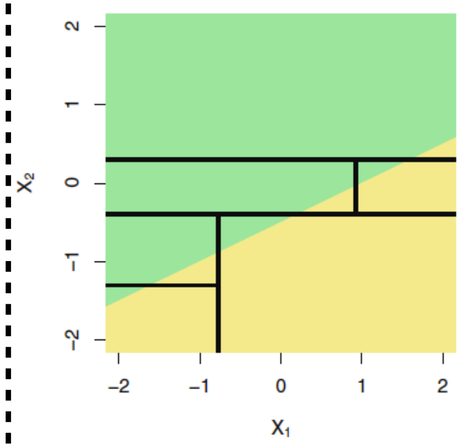
If true feature-response relationship matches model class our method “has in mind,” will take fewer parameters (less variance) to fit underlying trend (less bias).

Moral of the story: It’s good to know several prediction methods. Seek out the ones whose underlying model class you think matches the true feature-response relationship.

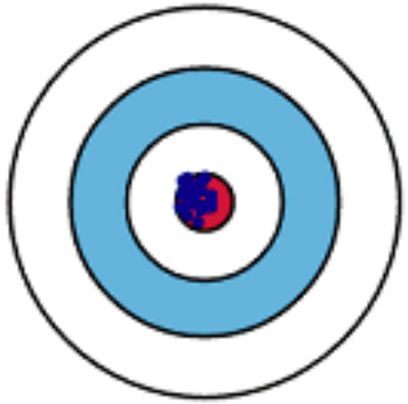
Right model class



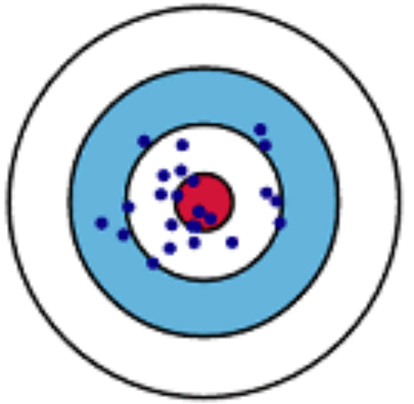
Wrong model class



Low Variance



High Variance



https://www.listendata.com/2017/02/bias-variance-tradeoff.html

Looking beyond STAT 471

What comes next?

- Where are data mining and machine learning going in the future?
- How do I learn more about data mining and machine learning?
- What other topics are relevant to data science beyond predictive modeling?
- What jobs out there value the skills I learned in STAT 471? How do I get those jobs?

Where are data mining and ML heading?

- Businesses increasingly data driven; data mining and ML will continue becoming increasingly common in marketing, finance, e-commerce, etc.
- Regression and tree-based methods will continue to be used for general-purpose data mining; deep learning for images and natural language.
- Datasets are becoming increasingly bigger, so more emphasis will be placed on large-scale/cloud computation and parallelization.
- As deep learning matures, more emphasis will be placed on understanding and interpretation, making it more safe, robust, and fair, developing theory.

Learning more about data mining and ML

Computation	Theory	Deep learning
• Python programming (CIS192, STAT477)	• Probability (STAT430, ESE301) • Linear algebra (MATH240, MATH312) • Calculus (MATH114, MATH115)	
• Large-scale computing (NETS212,CIS545) • Databases (CIS450)	• Mathy machine learning (CIS520,ESE545) • Optimization (STAT 481, CIS515, ESE504, ESE605)	• Deep learning (CIS522) • Computer vision (CIS680) • NLP (CIS530)

Offered next semester!

Beyond predictive modeling

Beyond predictive modeling

Not every problem in the world can be solved by machine learning!

Beyond predictive modeling

Not every problem in the world can be solved by machine learning!

Statistical inference versus prediction: understanding the world is different from predicting it. E.g., science is driven by understanding rather than prediction.

Beyond predictive modeling

Not every problem in the world can be solved by machine learning!

Statistical inference versus prediction: understanding the world is different from predicting it. E.g., science is driven by understanding rather than prediction.

Causal inference: Cause and effect is the ultimate question in science and policy. Decision makers (e.g. the FDA) need to estimate the effects of different policies.

Beyond predictive modeling

Not every problem in the world can be solved by machine learning!

Statistical inference versus prediction: understanding the world is different from predicting it. E.g., science is driven by understanding rather than prediction.

Causal inference: Cause and effect is the ultimate question in science and policy. Decision makers (e.g. the FDA) need to estimate the effects of different policies.

Quantifying uncertainty: If a neural network predicts there is no pedestrian ahead with probability 0.99, what does this mean? Can we make a statistical guarantee?

Beyond predictive modeling

Not every problem in the world can be solved by machine learning!

Statistical inference versus prediction: understanding the world is different from predicting it. E.g., science is driven by understanding rather than prediction.

Causal inference: Cause and effect is the ultimate question in science and policy. Decision makers (e.g. the FDA) need to estimate the effects of different policies.

Quantifying uncertainty: If a neural network predicts there is no pedestrian ahead with probability 0.99, what does this mean? Can we make a statistical guarantee?

Robustness and safety: Will the self-driving car recognize a pedestrian if she is holding an umbrella? In what sense can we ensure the robustness of an algorithm?

Beyond predictive modeling

Not every problem in the world can be solved by machine learning!

Statistical inference versus prediction: understanding the world is different from predicting it. E.g., science is driven by understanding rather than prediction.

Causal inference: Cause and effect is the ultimate question in science and policy. Decision makers (e.g. the FDA) need to estimate the effects of different policies.

Quantifying uncertainty: If a neural network predicts there is no pedestrian ahead with probability 0.99, what does this mean? Can we make a statistical guarantee?

Robustness and safety: Will the self-driving car recognize a pedestrian if she is holding an umbrella? In what sense can we ensure the robustness of an algorithm?

Fairness: In what sense can a prediction rule be considered fair? How can we assure that our predictive rules live up to this standard?

Data science jobs

Harvard Business Review

Data Scientist: The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data. by Thomas H. Davenport and D.J. Patil

From the Magazine (October 2012)

glassdoor

50 Best Jobs in America for 2021

Best Jobs | 2021 | United States

Share

	Job Title	Median Base Salary	Job Satisfaction	Job Openings	
#1	Java Developer	$90,830	4.2/5	10,103	View Jobs
#2	Data Scientist	$113,736	4.1/5	5,971	View Jobs
#3	Product Manager	$121,107	3.9/5	14,515	View Jobs
#4	Enterprise Architect	$131,361	4.0/5	10,069	View Jobs
#5	Devops Engineer	$110,003	4.0/5	6,904	View Jobs

What does a Data Scientist do?

Data scientists utilize their analytical, statistical, and programming skills to collect, analyze, and interpret large data sets. They then use this information to develop data-driven solutions to difficult business challenges. Data scientists commonly have a bachelor's degree in statistics, math, computer science, or economics. Data scientists have a wide range of technical...

Read More

Average Years of Experience

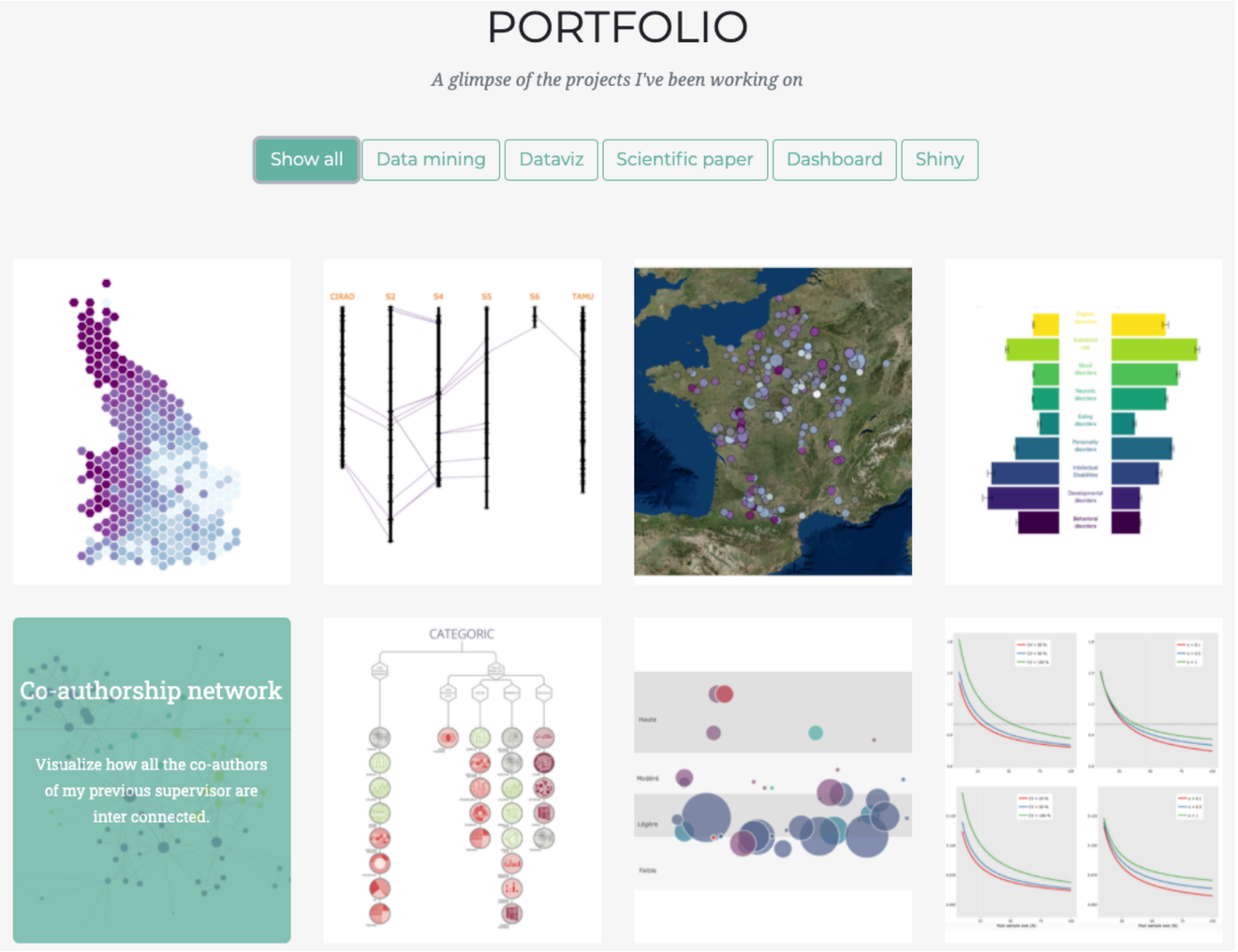
0 - 1	8%
2 - 4	80%
5 - 7	6%
8+	6%

Common Skill Sets

- Machine Learning
- Python
- Hadoop SPARK
- SQL
- Statistics
- Natural Language Processing
- Algorithms
- Programming Languages

How to get a data science job?

- Learn the skills through classes or on your own.
- Build your skills through data science projects.
- Share your work by posting code on Github and making a portfolio of your projects.
- Apply to internships to gain data science experience.



https://www.yan-holtz.com/