LECTURE 25

Conclusion

What you learned and didn't learn in this class. A farewell.

Agenda

- Logistics for the final.
- What did we teach you in this class?
- What’s next?
- Fun facts

Logistics

Final exam

The final exam is on **Friday, December 17th from 10:00-11:40**. It is worth 30% of your grade.

Important notes about the final:

- **Regex will not be on the final.**
- The exam is closed book closed note
- Please bring type two pencil with you
- No need for calculator

Feedback

Please fill out the course evaluation! (By 12/12)

As long as you submit proof that you filled out the evaluation on canvas, you will receive 1 bonus point!

What did you learn in this class?

What were we supposed to teach you?

Prepare

Prepare students for advanced courses in **data management**, **machine learning**, and **statistics**, by providing the necessary foundation and context.

Enable

Enable students to start careers as data scientists by providing experience working with **real-world data, tools, and techniques**.

Empower

Empower students to apply computational and inferential thinking to address **real-world problems**.

Note

- The following is a high-level overview of what we covered.
- You may find this useful in organizing your studying.
- But this is **not comprehensive!**

Data sampling and probability

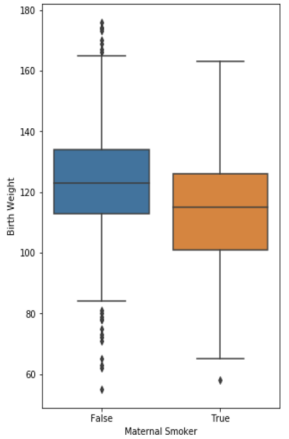
- Populations, samples, and sampling frames.
  - Sources of bias in sampling.
  - Types of samples (random samples vs. convenience samples, quota samples).
  - Benefits of random sampling.
    - Larger samples are not necessarily better.
    - Large biased samples can make things worse.
  - Sampling with and without replacement.
  - Binomial and multinomial probabilities.
- Random variables and distributions.
  - Expectation and its properties (e.g. linearity of expectation).
  - Variance and its properties.
  - Independence and correlation.
  - Sample means.

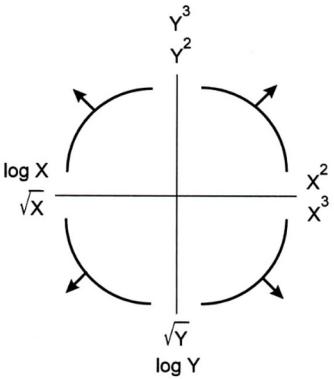
pandas, and data cleaning

- pandas as a means of working with tabular data in Python.
  - Series, DataFrames, and indexes.
  - .loc and .iloc.
  - Filtering, merging, grouping, pivoting, etc.
- Data cleaning and exploratory data analysis.
  - Structure, granularity, scope, temporality, faithfulness.
  - Handling missing values.

Regex and visualization

- Regular expressions as a means of identifying and extracting patterns in text.
  - Python string expressions.
  - regex101.com is your friend!
- Visualization.
  - Encodings and distributions.
  - When and how: bar plots, rug plots, histograms, density curves, box plots, violin plots, scatter plots, hex plots, contour plots.
  - Describing distributions in statistical terms (tails, skew, modes, outliers).
  - Principles of scale, conditioning, perception, and context.
  - Kernel density estimation.
  - Transformations.





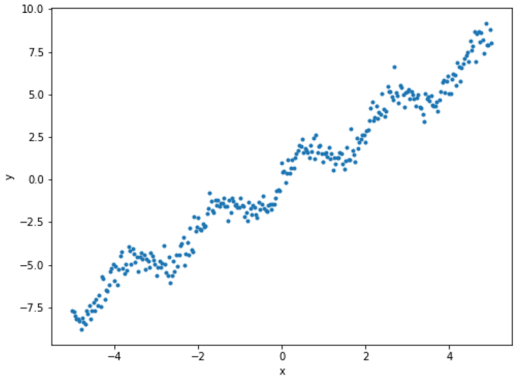
Modeling and linear regression

- The modeling “recipe”.
  - Choose a **model** (e.g. constant model, linear regression, logistic regression).
  - Choose a **loss function** (e.g. squared loss, absolute loss, cross-entropy loss).
  - Determine optimal parameters that **minimize** average loss (i.e. **empirical risk**) on training data.
- Simple linear regression.
  - Correlation coefficient r.
  - Finding optimal parameters by hand with calculus.
  - RMSE, Multiple R^2.

- Ordinary least squares.
  - **Regression**: given features, compute a real-valued prediction for each observation.
  - Design matrices, residuals, orthogonality.
  - Normal equation.

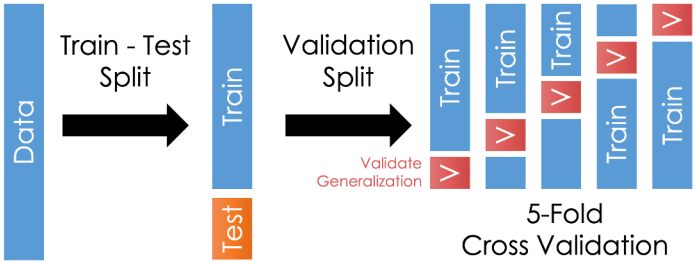
Feature engineering, bias-variance tradeoff

- Feature engineering: how do we model non-numeric and non-linear relationships, using linear models?
  - Categorical: one-hot encoding, bag-of-words, n-gram.
  - Polynomial features.
  - Overfitting to training data.
- Bias-Variance tradeoff
  - What are the underlying assumptions of randomness when modeling? (true model, random noise, etc.)
  - As we increase complexity, what happens to model bias? Model variance?
  - Model risk and its decomposition.

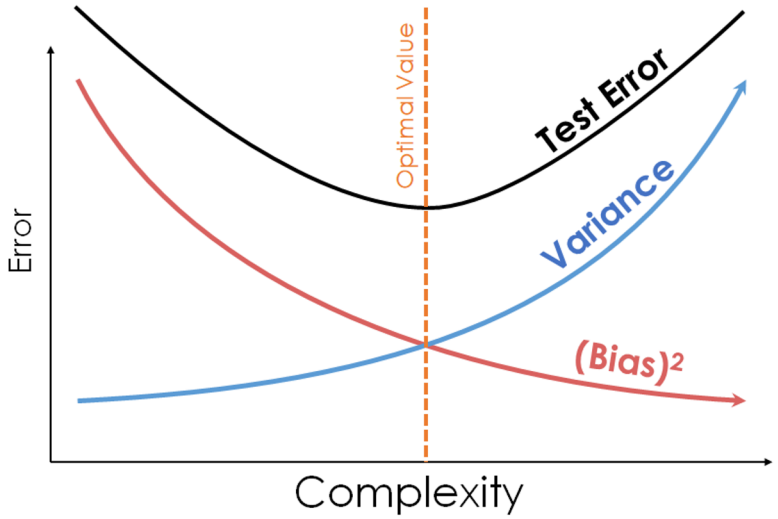


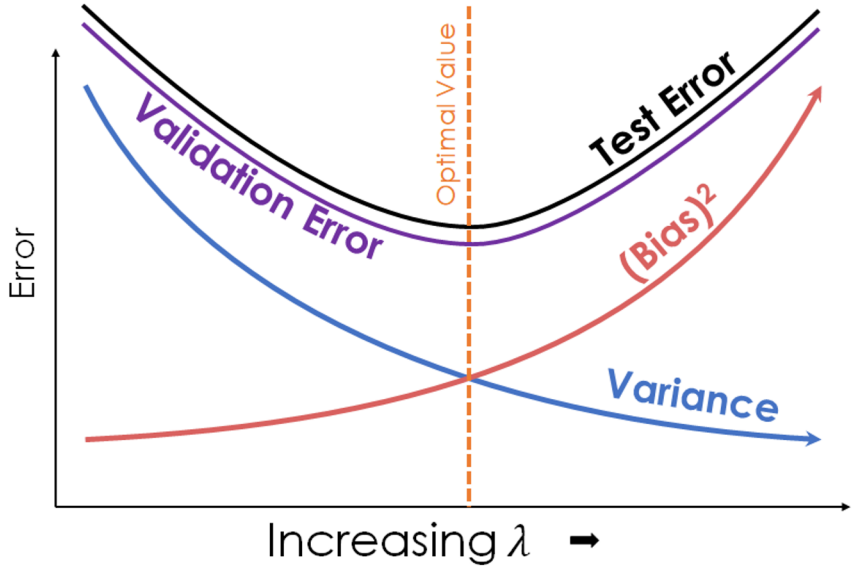
Regularization and cross-validation

- Regularization as a means of controlling model complexity.
  - Penalties on the norm of parameter vectors.
  - Ridge and LASSO.
  - Effects of the regularization hyperparameter on model complexity.
- Cross-validation as a means of selecting hyperparameters and/or sets of features.
  - Train-test splits, and why they're necessary.
  - Cross-validation as a means of estimating model performance on testing data using just training data.



Generalization

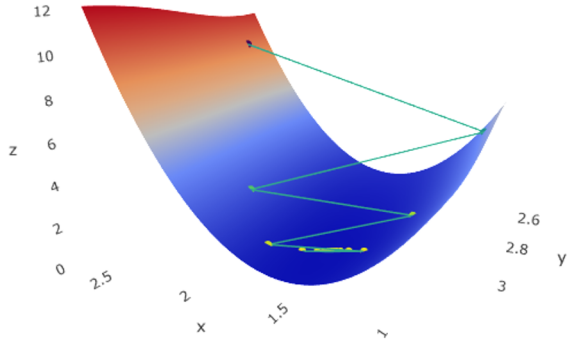


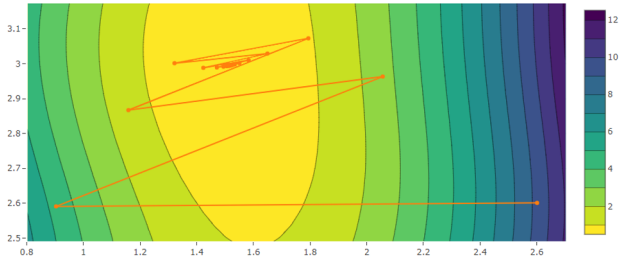


Our ultimate goal is to build models that generalize well to unseen data. The bias-variance tradeoff, training error, testing error, and cross-validation error all describe how well our model generalizes.

Gradient descent

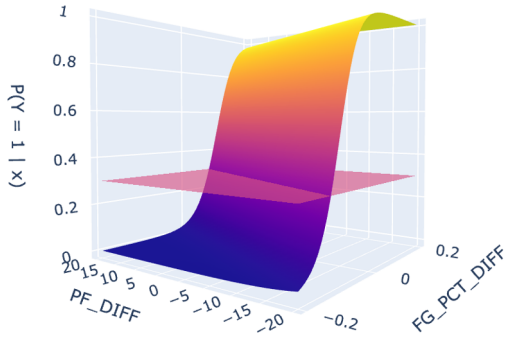
- Gradient descent as a means of numerically minimizing functions.
  - For our purposes, used to minimize average loss across a dataset.
  - Convexity.
    - If our loss function is not convex for our model, gradient descent may get stuck in a local minima.
  - Learning rates.
    - The size of the steps we take can affect whether or not we converge.
  - Stochastic gradient descent.

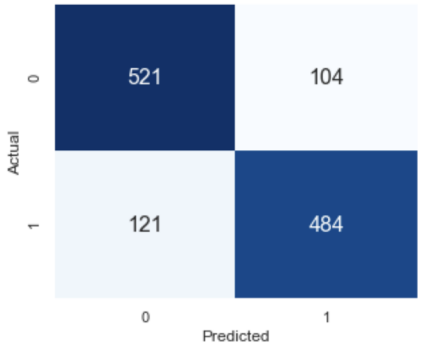




Logistic regression and classification

- Classification: given features, compute a discrete label for each observation.
- Logistic regression as a model of probabilities.
  - Linearity of log-odds.
  - The logistic function and its properties.
  - Cross-entropy loss vs. squared loss for logistic regression.
- Classification and classifier evaluation.
  - Thresholding and decision boundaries.
  - Linear separability, and the need for regularization.
  - Accuracy, precision, and recall. Confusion matrices.
  - PR curves, ROC curves, and AUC.

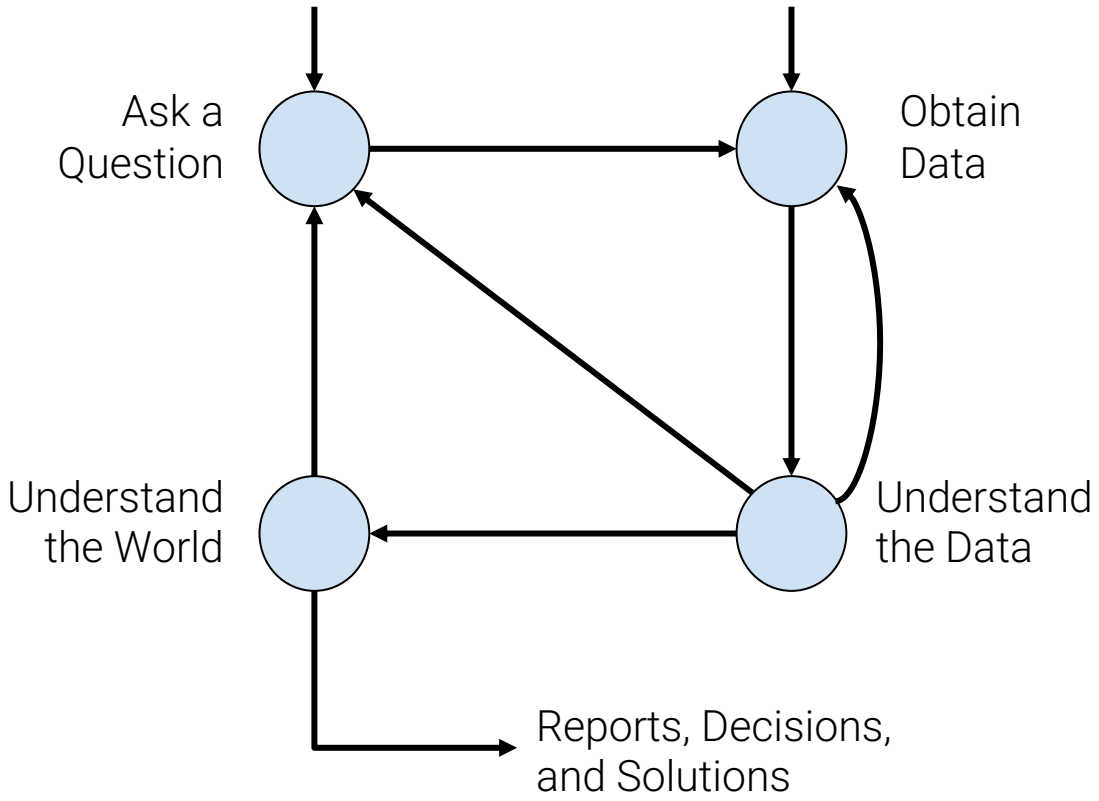




More models, model inference, PCA, clustering

- Other alternatives to regression and classification
  - Decision trees and random forests
  - Boosting: Adaboost, Realboos, Logitboost, XGBoost
  - SVM and SVR
- Model inference: using statistical tools to interpret our model's parameters.
  - Parameters, estimators, the bootstrap, and confidence intervals.
  - Multicollinearity.
- Unsupervised learning: instead of trying to learn the relationship between features and a response variable, we instead try and learn the pattern amongst the data itself.
  - Dimensionality reduction and PCA : How can we extract meaningful combinations of features from our high-dimensional data? What are "orthogonal basis vectors"?
  - Clustering: How can we segment our data into "groups" that are similar to one another?

Data science lifecycle



The data science lifecycle is a **high-level description** of the data science workflow.

Note the two distinct entry points!

Useful libraries and programming tools

What’s next

What didn’t we focus on this class?

- Causal inference.
  - How do we establish causality when we identify a correlation observed during EDA?
- Deep learning.
  - How can the machine do the hard work of picking the right features instead of requiring humans to pick them in advance?
- Decision making.
  - After we build our fancy regression/classification/clustering algorithms, what do we do next?
  - How do we avoid false discoveries?
- Time series analysis.
- Other flavors of machine learning (e.g. reinforcement learning).
- Open-ended exploration of problems and datasets picked by you.
- Non-tabular data (e.g. images/video, sensor-generated/spatial data, natural language...).

Real-world applications

A great way to strengthen your knowledge of the ideas you learned in this class (and to build a portfolio to help you find jobs!) is to use your new skills to analyze real-world data.

- There are countless sources of data available on the internet.
- Find one, load it into a notebook, and get to work!

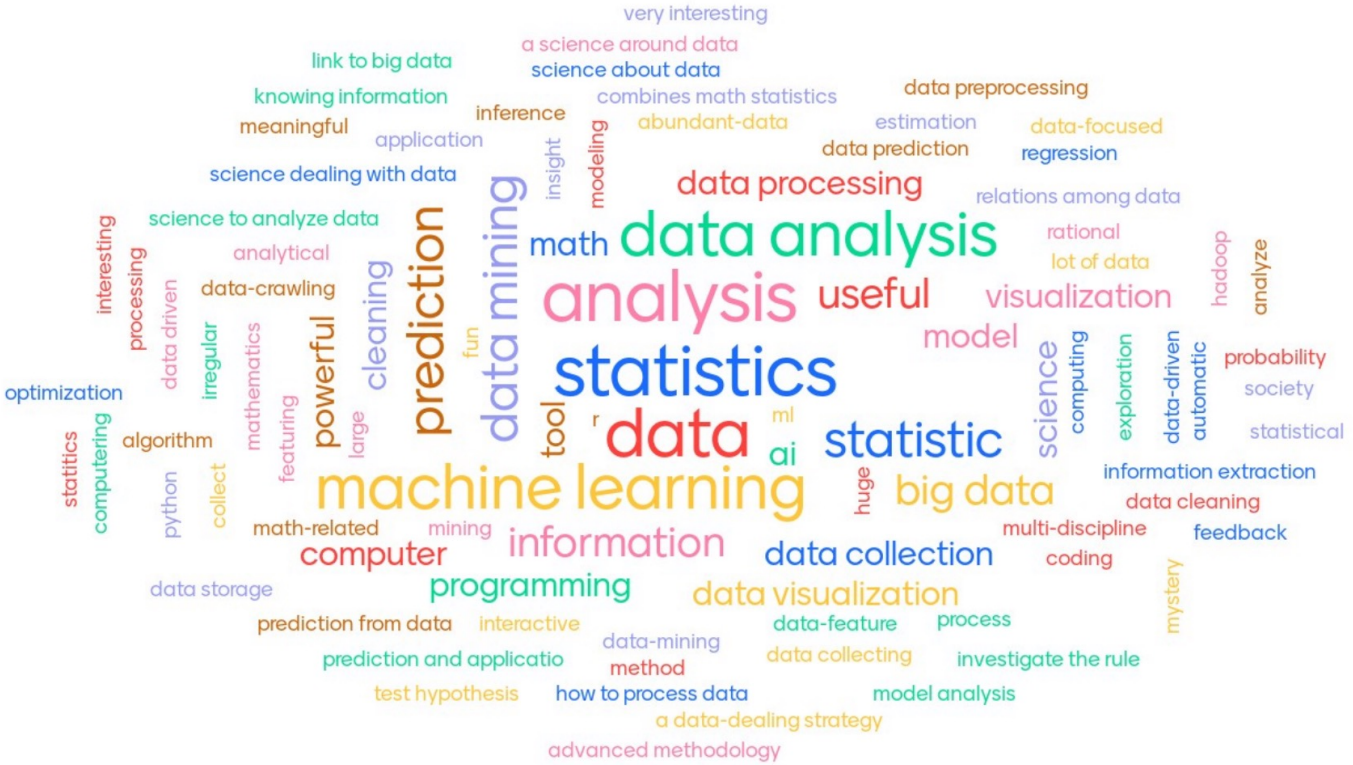
Places to look for data and applications of data science:

- Awesome Public Datasets GitHub repo.
- Kaggle.
- Towards Data Science.
- Five Thirty Eight.
- and so on!

Real-world applications

- Even though you're new to data science, you are also among the most skilled people in the world at data science. **Use your power wisely!**
- Let us know what you do next!

End (with some fun facts!)


very interesting
a science around data
link to big data
science about data
knowing information
combines math statistics
data preprocessing
meaningful
inference
abundant-data
estimation
data-focused
application
data prediction
regression
science dealing with data
insight
modeling
data processing
relations among data
science to analyze data
math
data analysis
rational
interesting
analytical
lot of data
processing
data-crawling
cleaning
prediction
data mining
analysis
useful
visualization
hadoop
analyze
data driven
irregular
fun
model
statistics
probability
optimization
mathematics
powerful
tool
computing
exploration
data-driven
automatic
society
large
r
data
ml
statistic
science
statistical
algorithm
featuring
ai
computering
statitics
machine learning
big data
information extraction
huge
data cleaning
collect
python
multi-discipline
feedback
math-related
mining
information
data collection
coding
computer
programming
data visualization
mystery
data storage
prediction from data
interactive
data-feature
process
data-mining
prediction and applicatio
data collecting
investigate the rule
method
test hypothesis
how to process data
model analysis
a data-dealing strategy
advanced methodology

Learning is about trial and error

Part 1

In [1790]:
```
ins['bid'] = ins['iid'].str.split("_").str[0]
ins['bid'] = ins['bid'].astype("int")
ins.head()
```

Out[1790]:

	iid	date	score	type	bid
0	100010_20190329	03/29/2019 12:00:00 AM	-1	New Construction	100010
1	100010_20190403	04/03/2019 12:00:00 AM	100	Routine - Unscheduled	100010
2	100017_20190417	04/17/2019 12:00:00 AM	-1	New Ownership	100017
3	100017_20190816	08/16/2019 12:00:00 AM	91	Routine - Unscheduled	100017
4	100017_20190826	08/26/2019 12:00:00 AM	-1	Reinspection/Followup	100017

Data Science is not only about understanding the world, but also understanding yourself!

Student	Dec 9 at 5:57am	07:15:14
TA	Dec 9 at 9:46pm	14:55:11
Student	Dec 9 at 12:22pm	03:09:33
Student	Dec 9 at 4:47pm	01:52:44
Student	Dec 9 at 7:28pm	05:14:23

Recruiting TA for 22 Spring!