Advanced Computing Techniques

Tim Head

3 April 2017

Welcome!

Timeline

- Six lectures
  - 3, 10, 24 April and 1, 8, 15 May
  - Always in this room, except 24 April (Batochime (Chemistry Building near BSP)))
- Six exercises
  - 1hr after each lecture.
  - We will discuss the questions from the previous week's lecture.
- One final project
  - Start on YYY, hand it in by XXX (to be decided)
  - Tim will propose topics for you to choose from, if you want to propose your own that is great, come see me afterwards.

Marks and credit

- Show that you learned something new!
  - It's personal.
- Grade:
  - *10% attendance*
    - Excuse yourself by email before 8.50am on the day.
  - *50% final project*
  - *20% handing in homework*
    - I will check: handed in? Does it run? Honest effort?
    - Did you make an effort and showed what you tried?
    - One of you leads discussion for each question during the exercise, at the end we will have the best possible answers.
    - Hand in via GitHub before the start of the lecture. Details on the exercise.
  - *20% participation*
    - Polite participation wins. This will be a team effort.

Technical things

- We will use python
  - We will use today's exercise to get you setup with what you need.
  - If you've never used python, don't worry.

- Use GitHub
  - Create an account and repository for the course if you don't already have one, we can help you during the exercise today.
  - Don't end up with: `final_project_v3_reviewed_final_iteration3.pdf`

Course overview

Machine-learning for scientists. Focus on applications and notation.

- **Introduction and jargon**
- Tree based methods
- Neural networks
- Measuring and predicting performance
- Dimensionality reduction
- Fine tuning
- Probabilistic data structures

Learning Goals

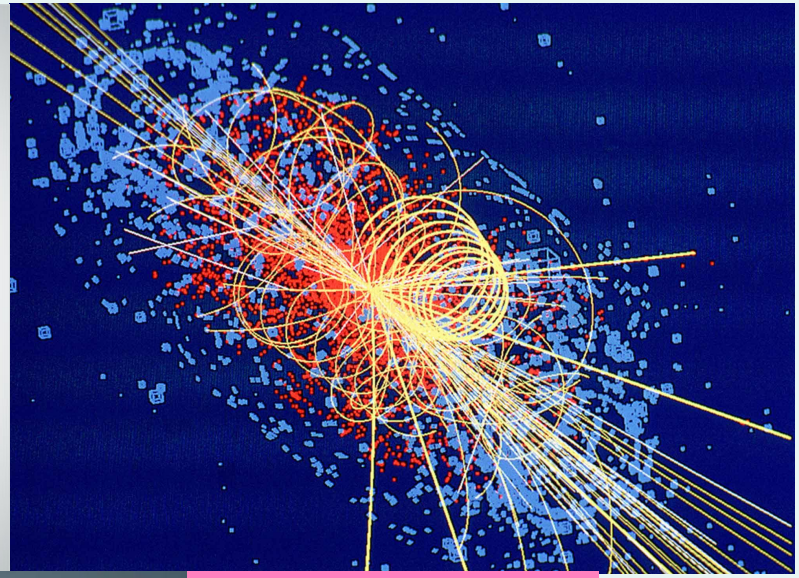
At the end of the course the student will be able to apply an ensemble of trees or neural network to a new data set. They will be able to justify and defend their choices of methods, cross-checks and optimisations.

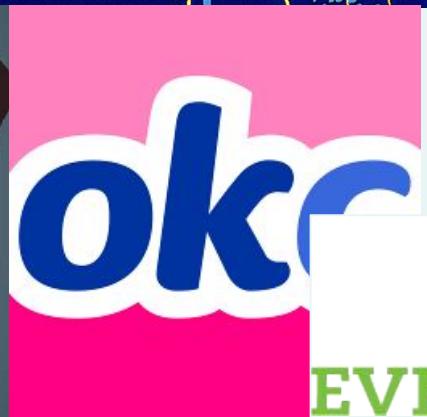
At the end of the course the student will be able to evaluate and give critical feedback on the choices of machine-learning techniques made in a paper related to the student's domain of expertise.

Where is machine-learning?

Spotify®

NETFLIX



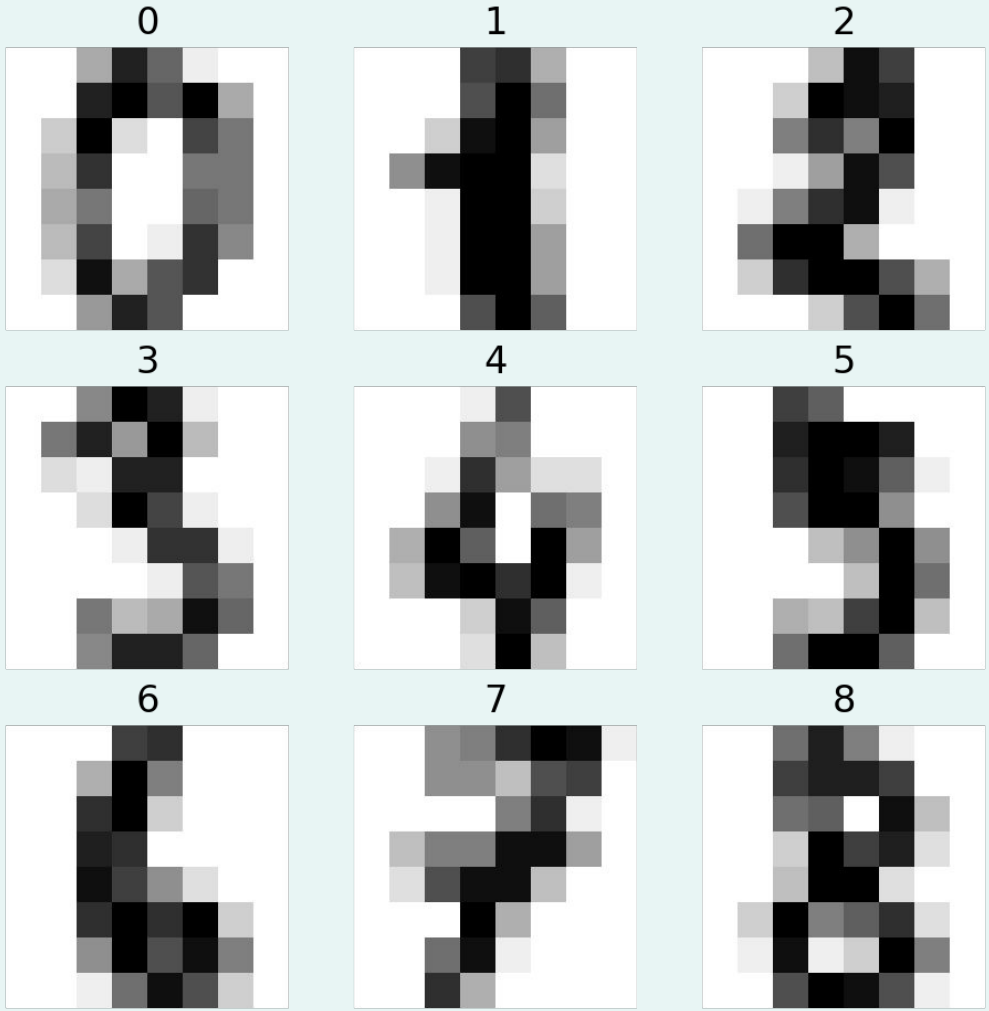


EVERNOTE

What is machine-learning?

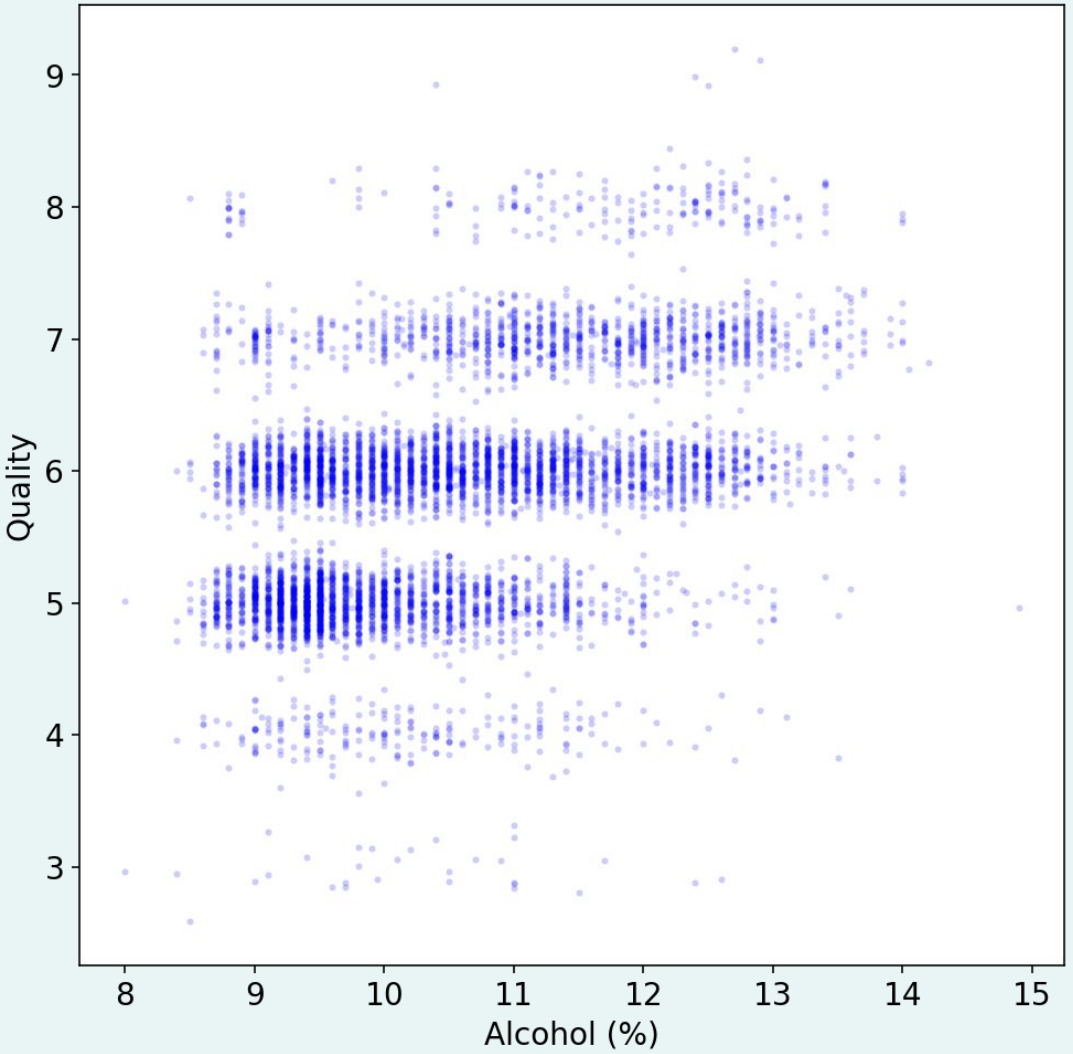
Digits

Recognise handwritten digits to read postcodes on envelopes.



Wines

The goal is to model wine quality based on physico-chemical tests.



... and more.

© Deepmind

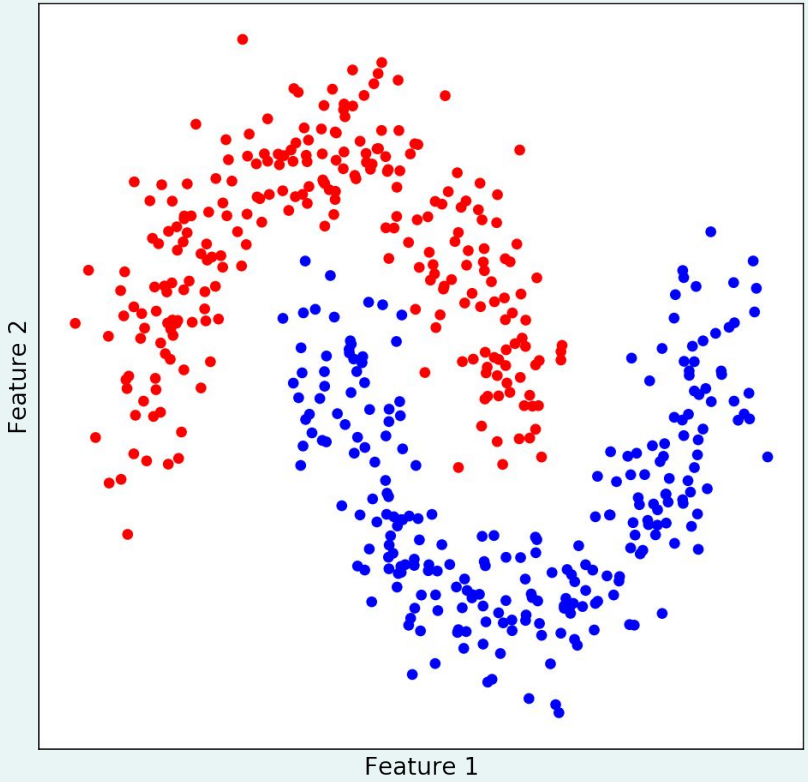
Machine learning

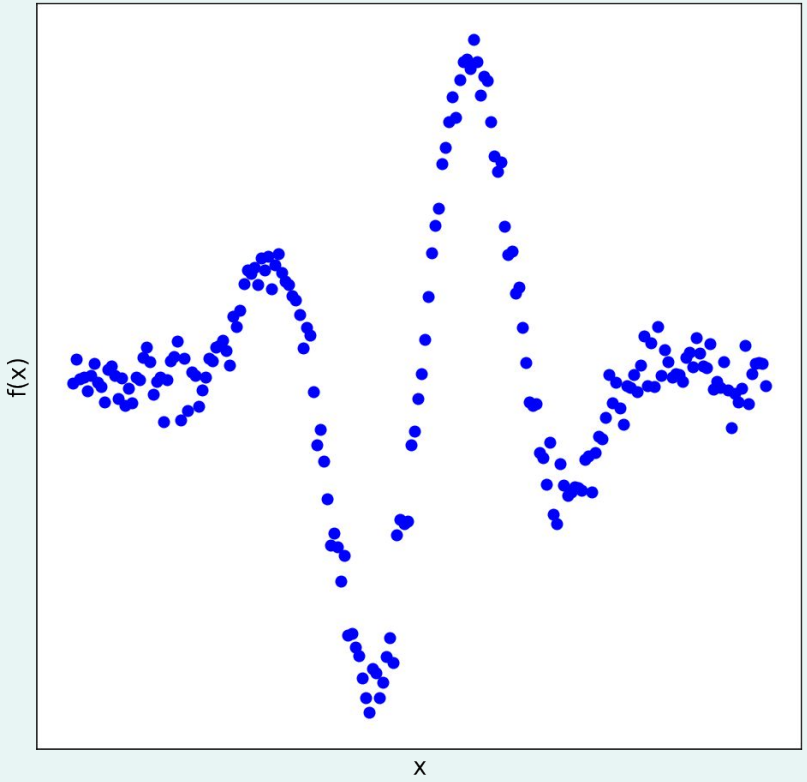
From Wikipedia, the free encyclopedia

For the journal, see Machine Learning (journal).

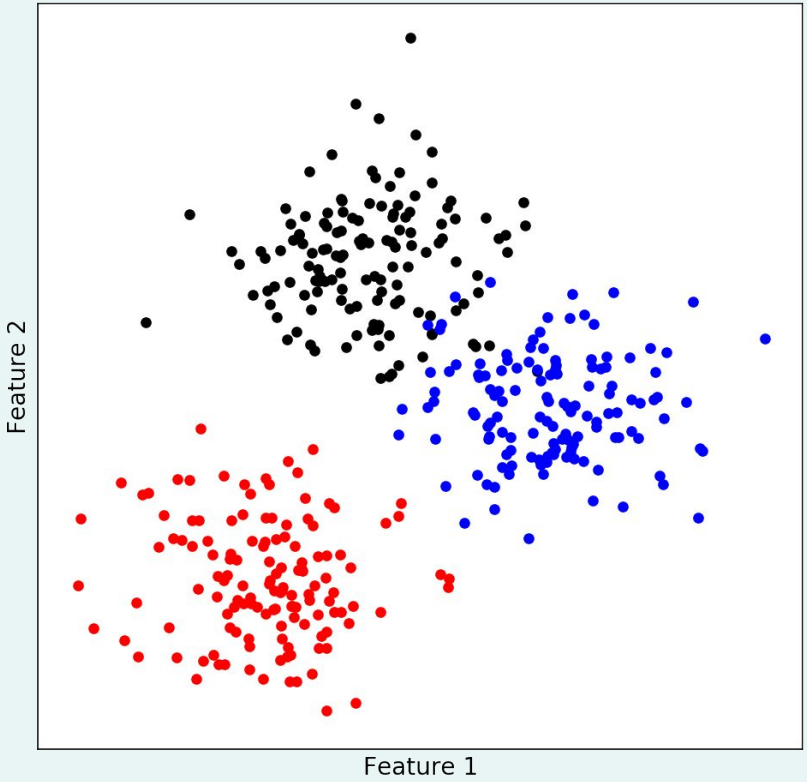
Machine learning is the subfield of computer science that, according to Arthur Samuel in 1959, gives "computers the ability to learn without being explicitly programmed."[1] Evolved from the study of pattern recognition and computational learning theory in artificial intelligence,[2] machine learning explores the study and construction of algorithms that can learn from and make predictions on data[3] – such algorithms overcome following strictly static program instructions by making data driven predictions or decisions,[4]:2 through building a model from sample inputs. Machine learning is employed in a range of computing tasks where designing

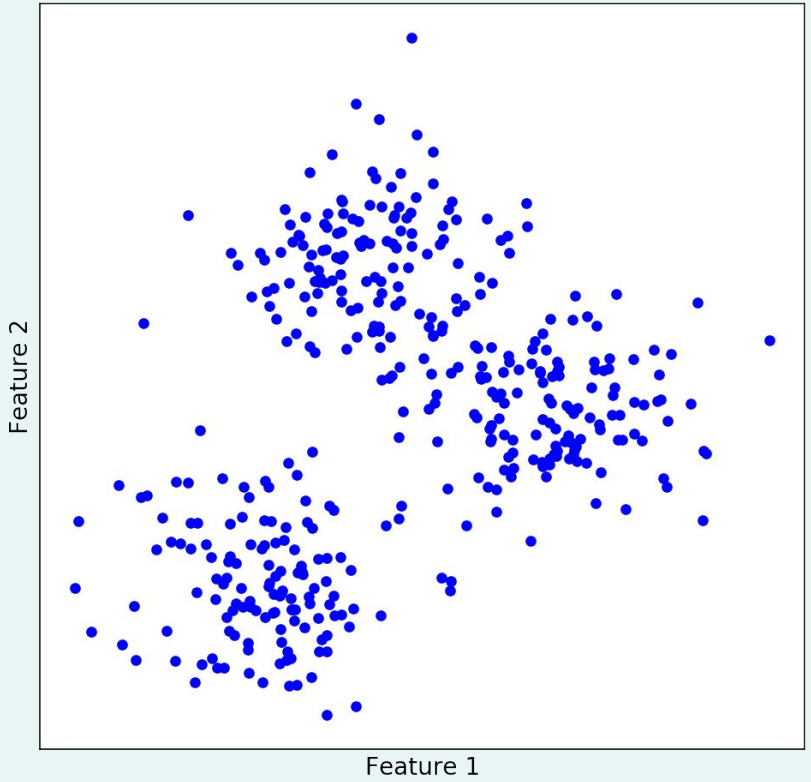
Regression vs Classification





Supervised vs Unsupervised





What is your data like?

A Classification - floats and integers
B Classification - images
C Regression
D Time series
E Unsupervised

http://bit.ly/adv-comp17-quiz

Live!

General problem statement

- Data comes as finite learning set $S = (\mathbf{X}, \mathbf{y})$
- Input samples are given as an array $\mathbf{X}$ of shape `(n_samples, n_features)`
  - Distance to EPFL, travel time, arrival time, ...
- Output values are given as an array $\mathbf{y}$ of shape `(n_samples,)`
  - Mode of transport

The goal is to build an estimator $f : \mathbf{X} \rightarrow \mathbf{y}$.

Linear regression

Live!

Recap

- Models have internal parameters which are “fit” to the training data.
- Parameters are optimised by minimising some loss function L.
  - Mean squared error for our linear model.
- All scikit-learn estimators follow the pattern of `est.fit(X, y)` and `est.predict(X)`

How good is the model?

Live!

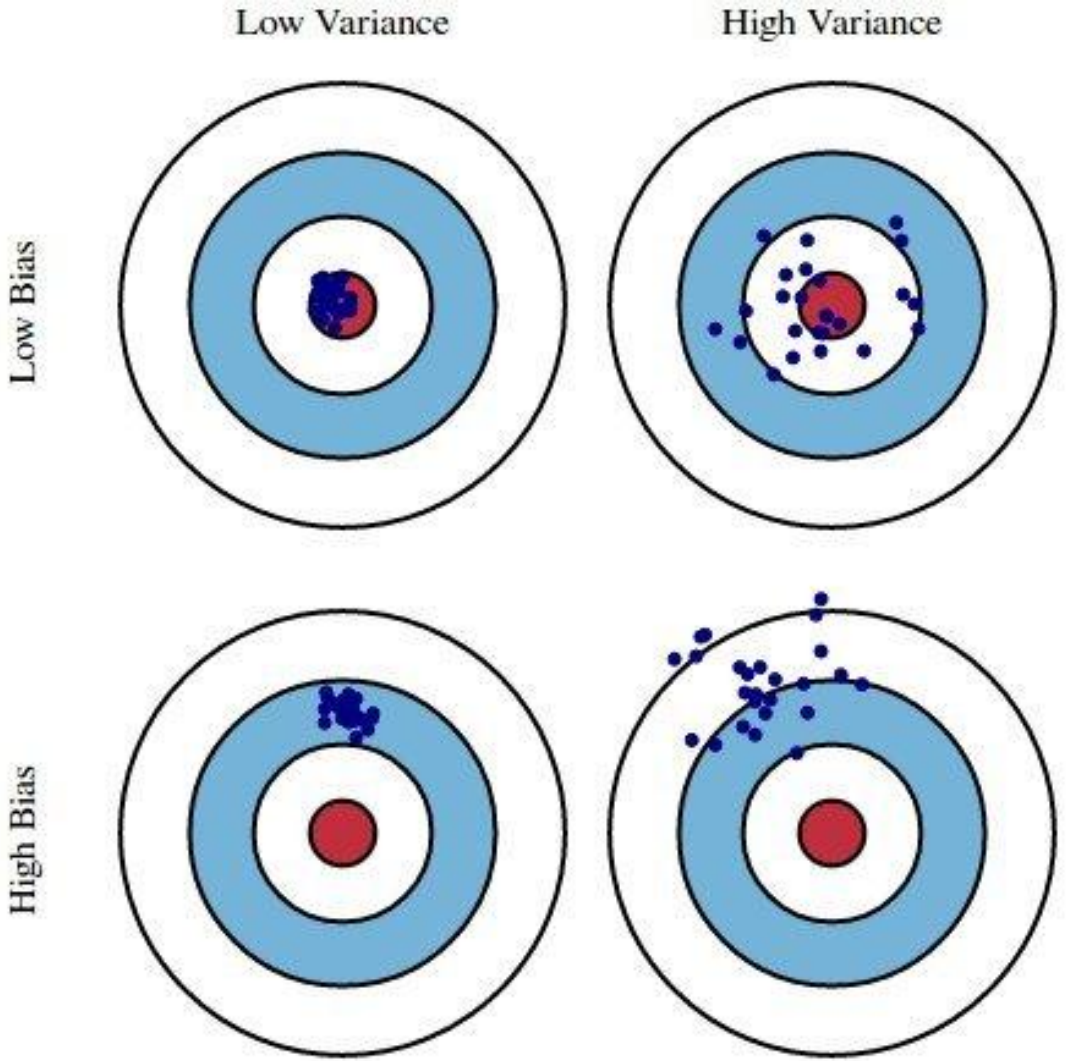
Recap

- To estimate the performance of your model use data that has been “locked away”.
- Almost all models try to minimise the training error when you fit them to the training dataset.
- The performance estimate from your training dataset will be biased.
- Training error will continually decrease as dataset size shrinks.

Model complexity

Live!

Bias vs variance



From “kdnuggets”.

Bias vs variance

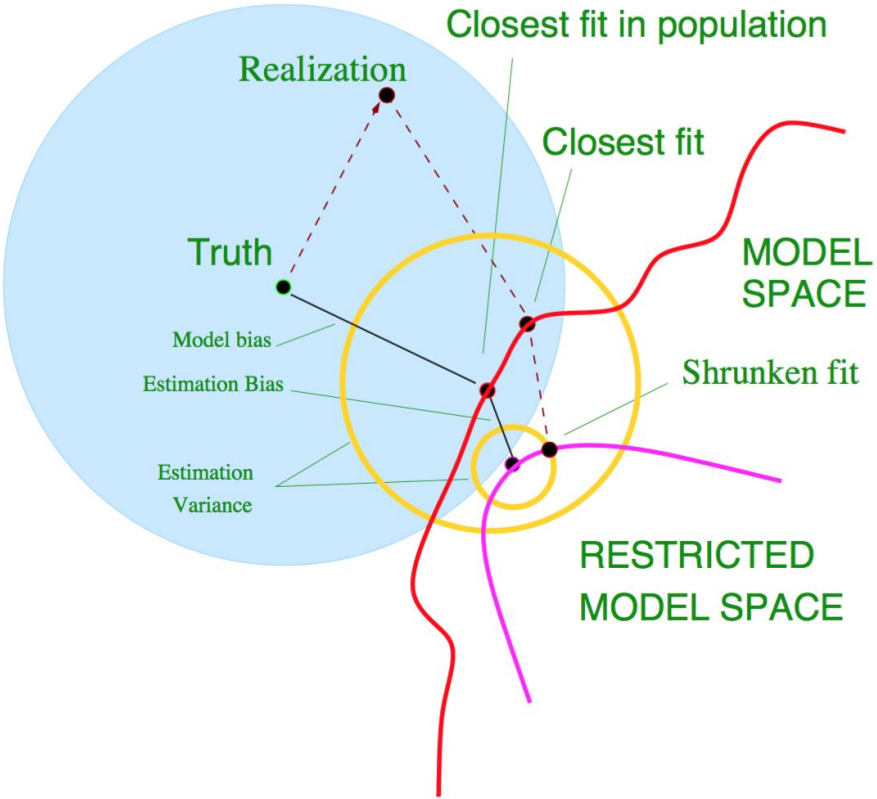


FIGURE 7.2. *Schematic of the behavior of bias and variance. The model space is the set of all possible predictions from the model, with the "closest fit" labeled with a black dot. The model bias from the truth is shown, along with the variance, indicated by the large yellow circle centered at the black dot labeled "closest fit in population." A shrunken or regularized fit is also shown, having additional estimation bias, but smaller prediction error due to its decreased variance.*

From "Elements of Statistical Learning".

Recap

- The training error continues to decrease as model complexity increases.
- Model complexity is controlled by hyper-parameters.
- Use an unseen dataset (test dataset) to find the sweet spot for your hyperparameters.
- Usually performance on the training data will be better than for the test dataset for all values of your hyper-parameter(s).
- Total error = bias^2 + variance + noise.
- Variance: spread of predictions fhat(x0) when fit on a new dataset.
- Bias: error introduced by simplifications made in the model.
- More flexible models will increase variance and decrease bias (rule of thumb).

Next week

- Tree based models and ensembles
- Neural networks

Next: let's get you setup!

...see you next week.