Overfitting and Capacity Control

John Klossner, *The New Yorker*

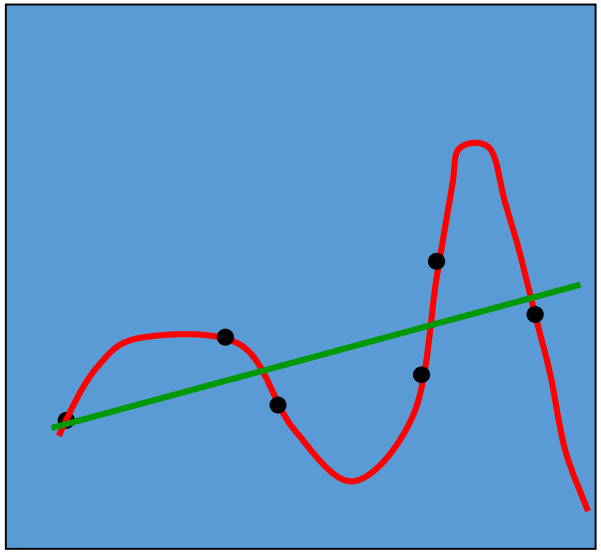
CSC411/2515: Machine Learning and Data Mining, Winter 2018

Michael Guerzhoy and Lisa Zhang

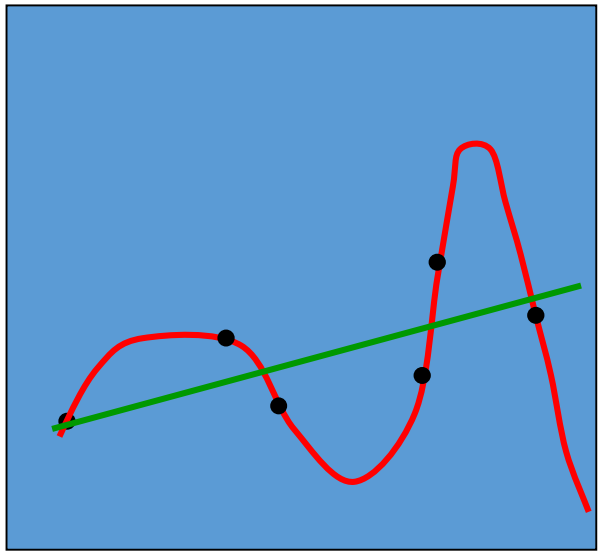
Slides from Geoffrey Hinton

Overfitting

- Overfitting happens when the model (e.g., a Neural Network, or k-NN, or…) models the specific training set rather than the underlying data from which the training set is taken
  - I.e., because the training set is too small, the network can do extremely well on the training set by modelling its peculiarities

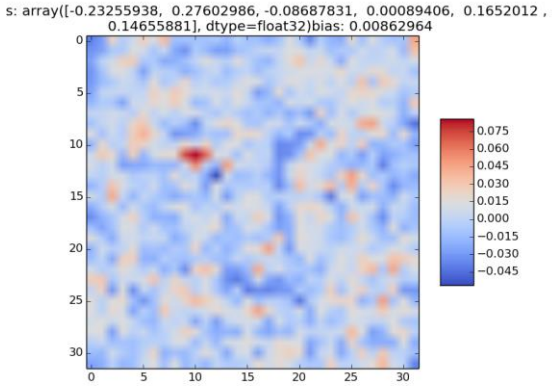


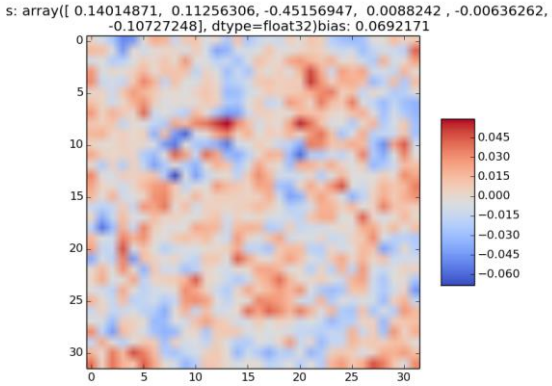
A Simple Example of Overfitting



- Which model do you believe?
  - The complicated model fits the data better.
  - But it is not economical
- A model is convincing when it fits a lot of data surprisingly well.
  - It is not surprising that a complicated model can fit a small amount of data.

Overfitting and Faces





- Above you see examples of W^0 that give near-100% performance on the training set
- The random spots you see are random regularities in the small training set being exploited – exploiting them on the test set won't work, and will possibly lead to bad performance

300 Hidden Units, 3000 epochs, no regularization, 40 examples per actor, 6 actors

Overfitting: Summary

- The training data contains information about the regularities in the mapping from input to output. But it also contains noise
  - The target values may be unreliable.
  - There is sampling error: there will be accidental regularities just because of the particular training cases that were chosen.
- When we fit the model, it cannot tell which regularities are real and which are caused by sampling error.
  - So it fits both kinds of regularity.
  - If the model is very flexible it can model the sampling error really well. This is a disaster.
- Overfitting: a model *making predictions based on accidental regularities in the training set*

How Overfitting Faces might Work

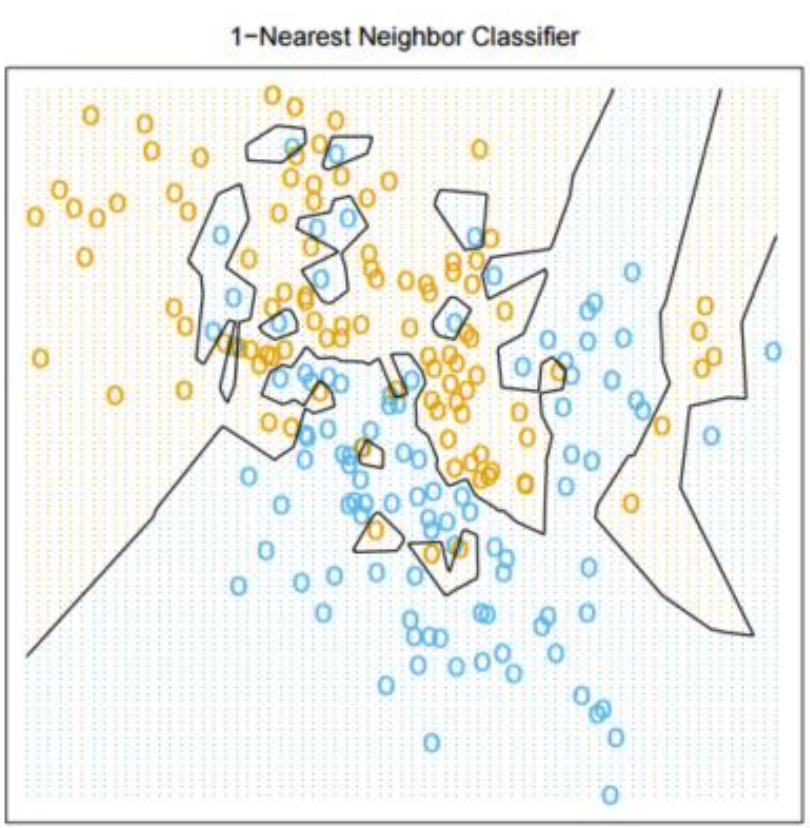
- See the “How Networks See” lecture!

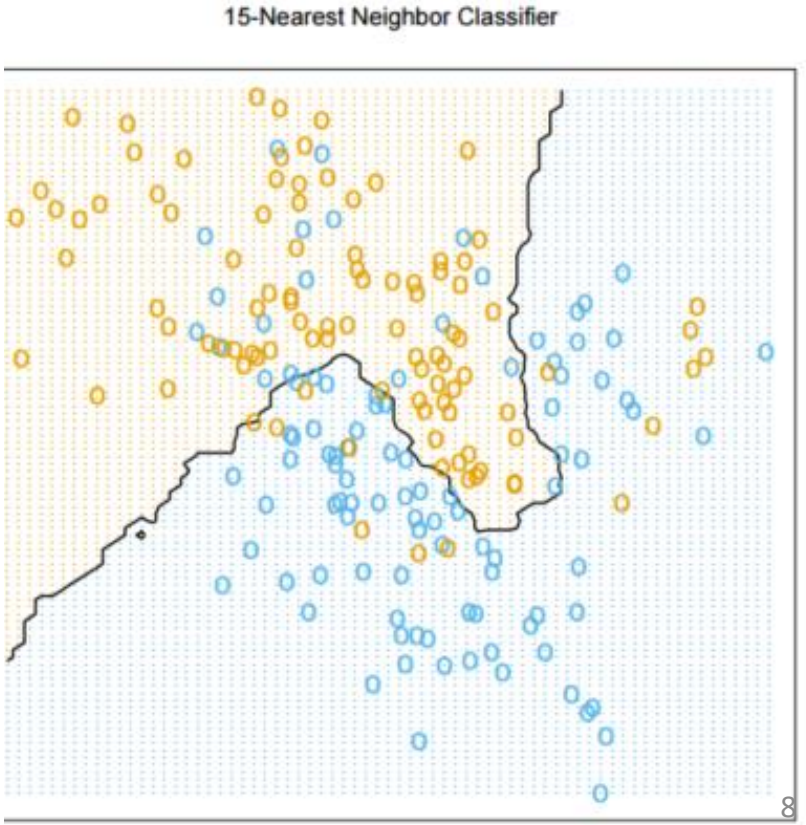
Preventing overfitting

- Use a model that has the right capacity:
  - Capacity: ability to produce different outputs depending on the input
    - Need enough to model the true regularities
    - Want to not have enough capacity to also model the spurious regularities (assuming they are weaker)
- Fitting curves in 2D:
  - Only fit lines, not higher-degree polynomials (example on the board)
  - Only fit quadratics, not higher degree polynomials

Reminder: Nearest Neighbours

- More nearest-neighbours ➡ less capacity
  - More complicated decision surfaces are not possible





Limiting the Capacity of a Neural Network

- Limit the number of hidden units
- Limit the size of the weights
- Stop the learning before it has time to overfit
- Dropout