Machine Learning for the Social Sciences

Intro, Review of key concepts

Andrew Peterson

10.04.2018

Université de Genève

ML Concepts: Warm-up /Review

What are the elements of a ML-algorithm?

Elements

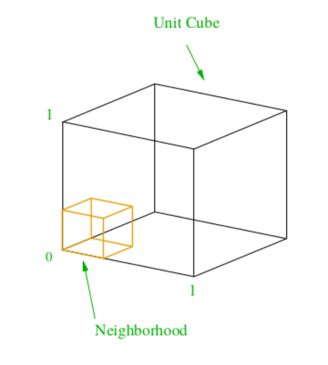
- cost function (can relate to real-world costs...)
- function class
- regularization penalty

  (implicit in above: assumptions/ restrictions - linearity, local smoothness, etc...more later)

What is the curse of dimensionality?

Curse of dimensionality (KNN example)

- p-dimensional hypercube, uniformly distributed points



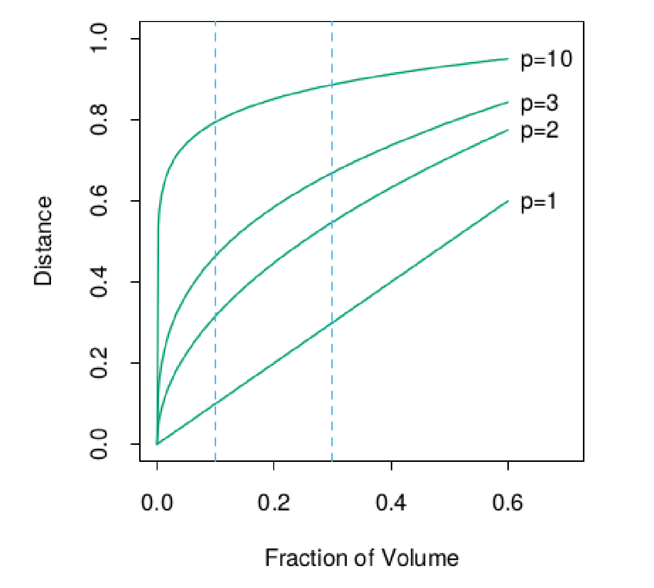
Source: Hastie et al. ESL

- What size box needed to capture a fraction of all observations, (e.g. to estimate local treatment effect), say $r = 0.1$?

Curse of dimensionality

- Need edge length: $e_p(r) = r^{\frac{1}{p}}$
- $p = 10$: need 63% of the range of each input variable to capture 1% data

Curse of dimensionality



Source: Hastie et al. ESL

Curse of dimensionality

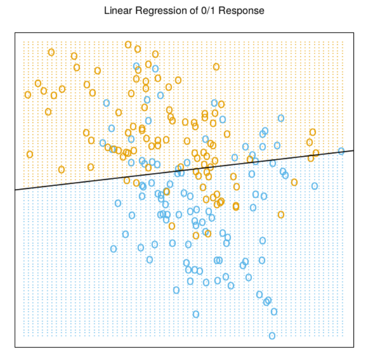
- Ex 2 (ESL): p-dim unit ball.
- Mean distance to nearest point: In high-dimensions, most data points are closer to edge than another point

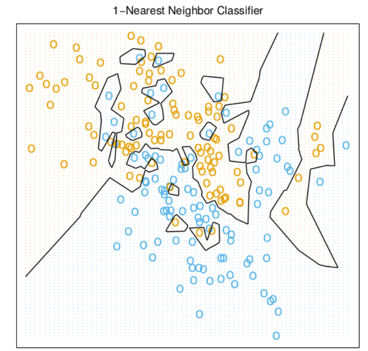
  $\rightarrow$ High dimensions not always intuitive
- Implications - e.g. matching, robustness, etc.

What is the bias-variance tradeoff?

How do we balance?

Bias-Variance





Source: Hastie et al. ESL

Bias-Variance: solution

- Minimize _?_ on cross-validation

  Expected prediction error

- (bad) example:

  step 1: select relevant words from all texts based on correlation with outcome

  step 2: Generate predictive model among relevant words using cross-validation

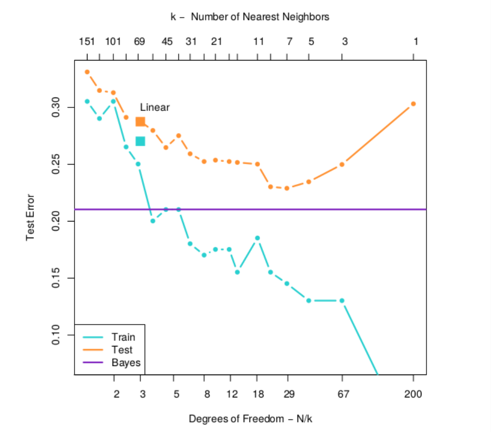
- NB: validation data $\neq$ test data!

Again, cost function should reflect use: e.g. might care more about false negatives for disease testing, etc.

What is overfitting?

Overfitting

- Overfitting is when there are too many parameters so that there is low bias but high variance.



Source: Hastie et al. ESL

Overfitting

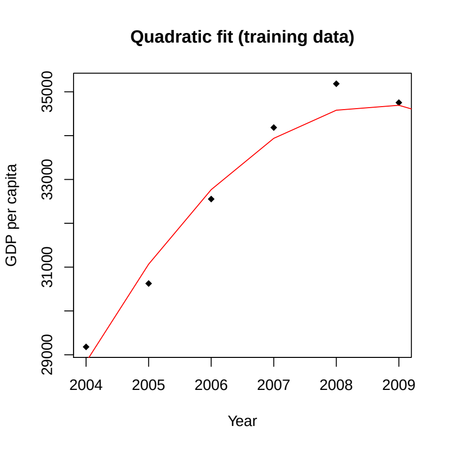
- ~~Overfitting is when there are too many parameters so that there is low bias but high variance.~~
- Overfitting is when the model fails to generalize to the test set. (One way to see the difference: if allow observations to have different error structures - more parameters but not higher variance...)
- Some models can “memorize” the training set (e.g. overtraining DNNs...)
- So although limiting model complexity helps prevent overfitting, they are non synonymous.
- For definitions of model complexity (effective number of parameters, VC-dimension, see ESL ch.7).

Overfitting

- Can arise when the criteria for model selection are not our ultimate evaluation
- e.g. "conceptual drift", Google Flu example, etc.
- Assumption: functional relationship in validation data same as that in test data.

Overfitting: Another example

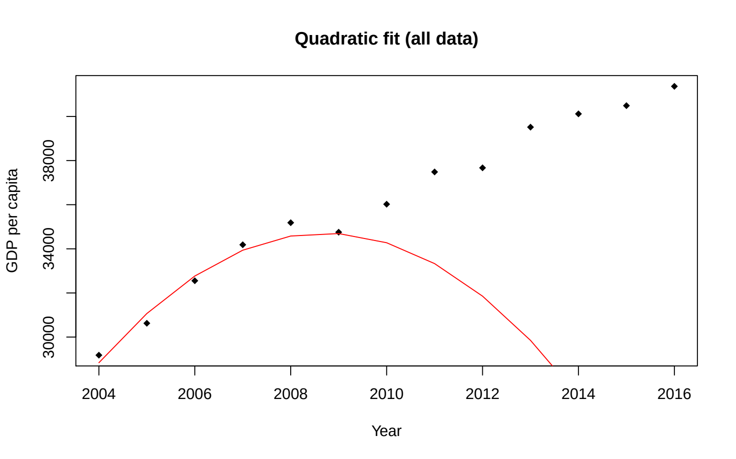
Predict values of French GDP/capita, 3 parameters, training data: 2004-2009



(Data from data.oecd.org)

Overfitting: Another example

Predict values of French GDP/capita, training data: 2004-2009



(Data from data.oecd.org)

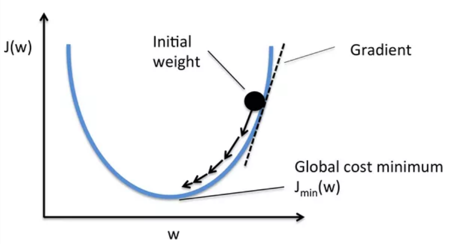
Overfitting: Solutions

- complexity penalty
- limit training
- training algorithms that prevent overfitting (e.g. SGD, NN dropout, ...)

each of these has consequences, as we will see...

Aside: Stochastic Gradient Descent

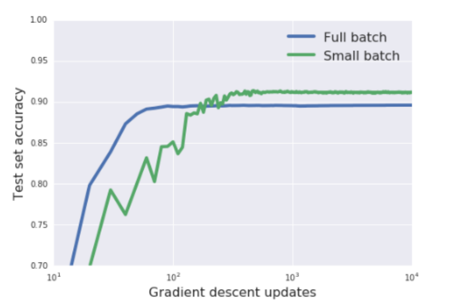
- Estimate gradient from small batch of observations
- Each step updates parameters towards min cost (hopefully)



- Athey undersells - is more than just computationally efficient

Stochastic Gradient Descent

- noisy but unbiased steps
- may have benefits: small batches may prevent overfitting (like wearing big boots on a golf course)



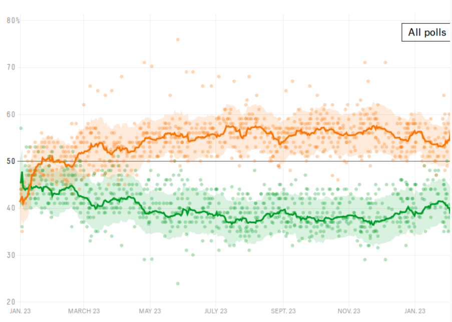
Source: Smith & Le, “A Bayesian Perspective on Generalization...”

Stochastic Gradient Descent

- So - we can be concerned about overfitting as a result of failing to find global minima in our optimization process, not (just) because there are too many parameters.

Random question

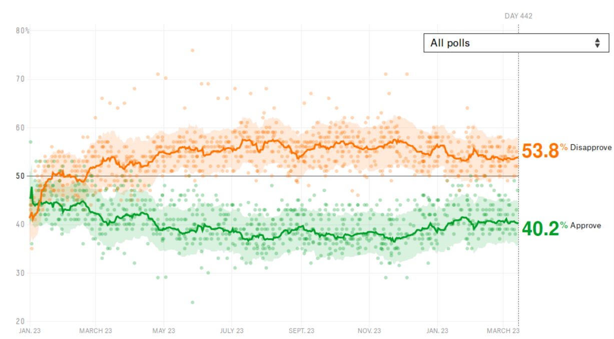
- Trump approval ratings especially low. What's the best prediction? (same? higher? lower?)



Source: FiveThirtyEight.com

Shrinkage

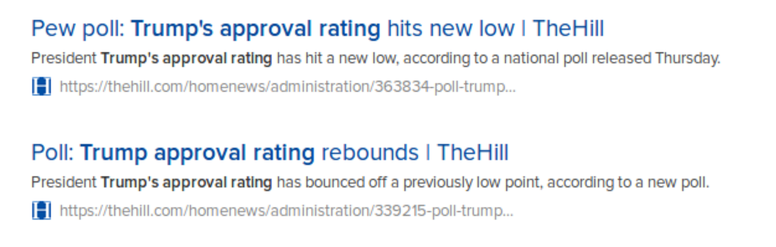
- Shrinkage: e.g. regression - r^2 on new data ↓
- Can do better by shrinking (coef) estimates towards zero (James-Stein)



Source: FiveThirtyEight.com

Shrinkage

- anecdotally (internet search results while looking for data to illustrate this point):



Pew poll: **Trump's approval rating** hits new low | TheHill

President **Trump's approval rating** has hit a new low, according to a national poll released Thursday.

https://thehill.com/homenews/administration/363834-poll-trump...

Poll: **Trump approval rating** rebounds | TheHill

President **Trump's approval rating** has bounced off a previously low point, according to a new poll.

https://thehill.com/homenews/administration/339215-poll-trump...

- So what? Suggests another aspect of the problem of generalization error.