More Machine Learning

William Lowe

Hertie School

4th October 2020

Moar ml

Plan

- → Trouble in high dimensions
- → Classification
- → Evaluating classifiers
- → Two ways to find a decision boundary
- → Forests, causal and otherwise

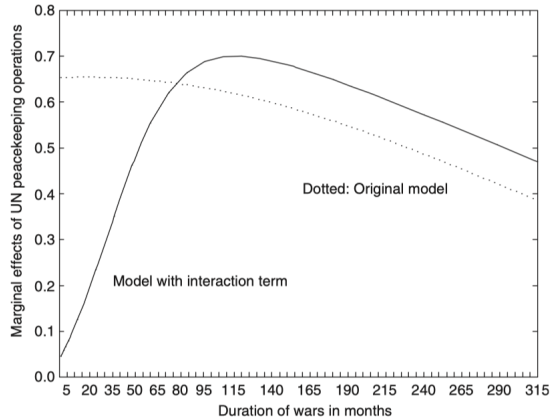
Dimension trouble

What's the problem with having a high dimensional problem?

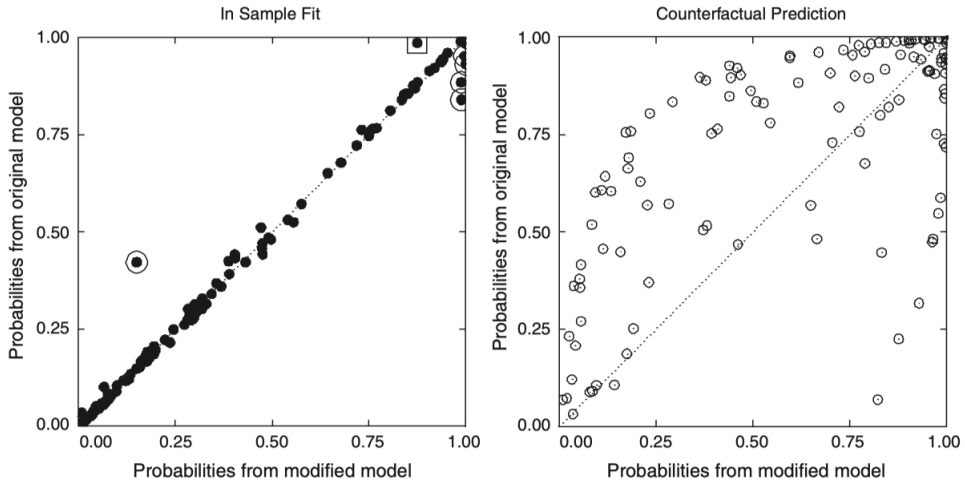
- → Sensitivity to model choices. That's 'variance' from last week)
- → Variation from model choices may (in general, will) only show up in the *counter*factuals

Example: The effectiveness of multilateral UN operations in civil wars (Doyle & Sambanis, 2000).

Reexamined by King and Zeng (2007).



Differing in the counterfactuals



More dimension trouble

What's the problem with having a high dimensional problem?

→ Specifically, having lots of confounders you want to control for

We've seen the extreme case:

→ more variables than cases: breaks regular models

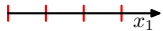
and the 'solutions'

→ assumptions, e.g. additivity
→ constraints, e.g. parameter regularization: $\beta_1 \dots \beta_K \sim \text{Normal}(0, \sigma^2)$

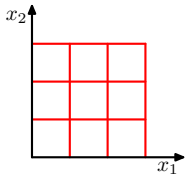
but let's consider the dimension issue directly

Dimensions

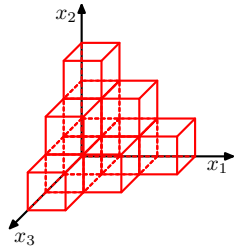
High dimensional space is very, very empty



$D = 1$



$D = 2$



$D = 3$

It's unlikely your observations can keep filling each cell of covariate value combinations as you keep deciding to measure more stuff

→ Note: this is one good motivation for balancing scores like the propensity score

Dimensions

High dimensional space is very, very weird

Consider our old friend, the Normal distribution. We'll take a concrete example:

- → Policy preferences, for D policies
- → The kind of thing you ask in a survey with $\geq D$ questions

Assume for a moment that the population has jointly normally distributed preferences

- → On any dimension, the probability of responses is Normal centered around the middle of the policy space
- → Policy preferences are uncorrelated (and therefore independent)

(Nothing much depends on these assumptions for making the point ahead)

Dimensions

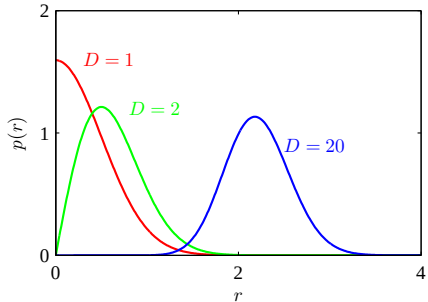
What sorts of *preference profiles* do we expect to see in a random sample?

Intuition says:

→ lots of 'centrists'

But it turns out, that depends on D

The 'typical set'



→ Note: the probability *density* always looks like $D = 1$ but it's spread out across a space of increasingly large D, so the mass diverges

→ This phenomenon is called the concentration of measure (link to the math)

At $D = 20$ centrists are (surprisingly) rare

Example: ideology and sophistication

Broockman (2016) points out that *sophisticated* respondents have low-D preferences

- → Their views on different policies are *not* independent
- → Equivalently, their effective D is lower than the number of questions you ask them

Unsophisticated voters have nearly independent preferences

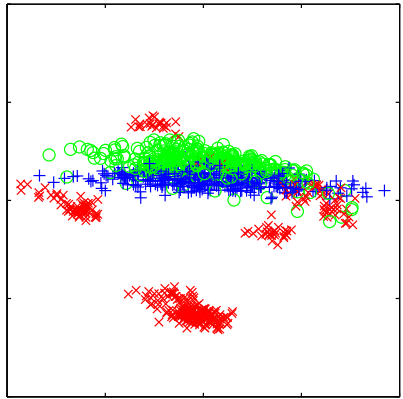
- → Many preference inference methods assume sophistication, e.g. averages of directed responses, scaling models, etc.
- → These will put sophisticated voters in the right place, and unsophisticated voters *in the middle anyway* despite them being mostly elsewhere

Ideology is a regularizer / dimensionality reducer / preference structurer…

Interlude: Ideology and representation

Simple (non-preference) data
How to summarize high D (here 2) in low D? (here 1)

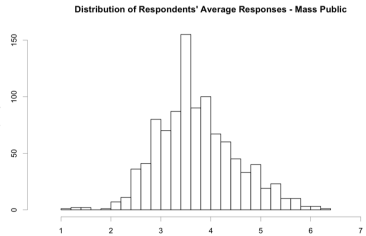
- → The most variation is on the x-axis
- → But significant variation on y!
- → …which will be collapsed into the middle of the left right axis as we lower D with, e.g. PCA

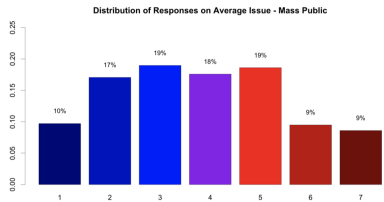


Interlude: Ideology and representation

Two ways to measure extremity of policy preferences

- → IRT scaling model (think: PCA or just a fancy index)
- → Averaging views issue by issue, not altogether





The curse of dimensionality

This general problem of empty high-D covariate space is called the 'curse of dimensionality'

Happily

→ Assertion: most real data relationships live in a subspace of the covariates ('sophistication' is widespread, because the social world is very not random)

However,

→ There is no guarantee this assertion is true
→ Even if it is, the structure of $X_1 \dots X_K$ may not be pleasantly linear, or additive
→ So $Y \longleftarrow X_1 \dots X_K$ may not be either

Assumptions and constraints from ML models

→ Smoothness: loess, splines, neural networks
→ Locality: k-nearest neighbours, kernel methods, trees

Dimension solutions

Interestingly sometimes deliberately *increasing* the dimensionality of a problem can help, e.g

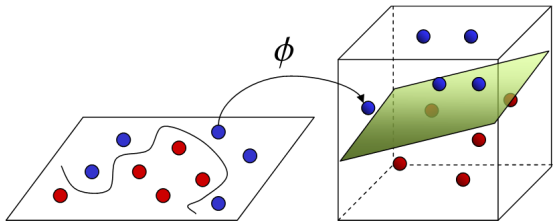
- → adding polynomials, logs, exps, etc. of covariates
- → this is the *feature space* (an expansion of the covariate space)

Dimension solutions

Interestingly sometimes deliberately *increasing* the dimensionality of a problem can help, e.g

- → adding polynomials, logs, exps, etc. of covariates
- → this is the *feature space* (an expansion of the covariate space)

Consider a classification problem: distinguishing Y = red vs Y = blue



Lots of things become *linearly separable* in a larger space, as e.g. leveraged by Support Vector Machines (SVMs)

Classification

We've been assuming a regression context for out ML so far, but we can also think about classification

Reminder: classification is two things, often confused. In a simple two class (0/1) classification

- → Estimating $E(Y \mid X_1 \dots X_K) = P(Y = 1 \mid X_1 \dots X_K)$
- → Deciding 1 or 0 in the light of $P(Y = 1 \mid X_1 \dots X_K)$

Implicitly you may be used to deciding 1 if $P(Y = 1 \mid X_1 \dots X_K) > 0.5$

However, it is often more costly to mistake a 1 for a 0 than a 0 for a 1, e.g.

- → 1 means a state will collapse in the next year (e.g. King & Zeng, 2001)
- → The losses are far from equal
- → Intuitively we should require lower probability to choose 1 when mistaking 1 for 0 is very costly

Classification

Decision theory:

- → L_{ij} is the cost of mistaking i for j e.g. L_{10} is the cost of mistaking a 1 for a 0
- → Minimize the expected L by choosing the i that minimizes

$$\sum_j L_{ij} P(Y = i \mid X_1 \dots X_K)$$

For 1/0 decisions another way to put this is in terms of a cutoff: Choose

$$\hat{Y} = \begin{cases} 1 & \text{if } P(Y = 1 \mid X_1 \dots X_K) > \frac{1}{1+C} \\ 0 & \text{otherwise} \end{cases}$$

where

$$C = \frac{L_{10}}{L_{01}}$$

Classification errors

From the loss function we can also identify two sorts of error

- → Mistaking a 1 for a 0: $P(\hat{Y} = 0 \mid Y = 1)$
- → Mistaking a 0 for a 1: $P(\hat{Y} = 1 \mid Y = 0)$

A useful and closely related pair of quantities are

$$P(\hat{Y} = 1 \mid Y = 1) = 1 - P(\hat{Y} = 0 \mid Y = 1) \qquad \text{(recall)}$$

$$P(Y = 1 \mid \hat{Y} = 1) = \frac{P(\hat{Y} = 1 \mid Y = 1)P(Y = 1)}{P(\hat{Y} = 1)} \qquad \text{(precision)}$$

Varying C expresses a tradeoff between these too

- → High C lowers the cutoff, which increases recall but decreases precision
- → Low C raises the cutoff which increases precision but decreases recall

Unknown losses, unknown tradeoffs

Sometimes we don't have (or can't commit to) some loss matrix L or a prefered balance between precision and recall

However, since each value of C implies such a loss / balance, we can ask how well a classifier does for *all possible* cutoffs

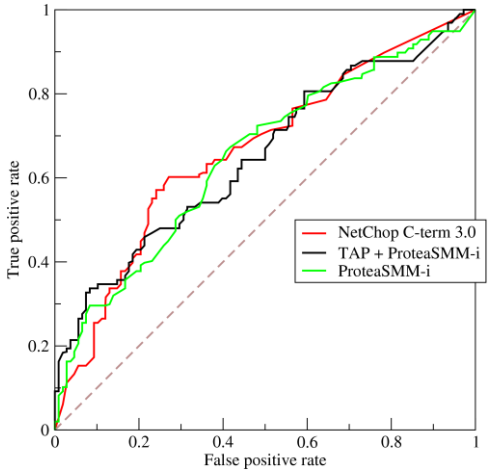
Traditionally we plot precision and recall in a *Receiver Operating Characteristic* (ROC) curve for a wide range of cutoffs

Warning:

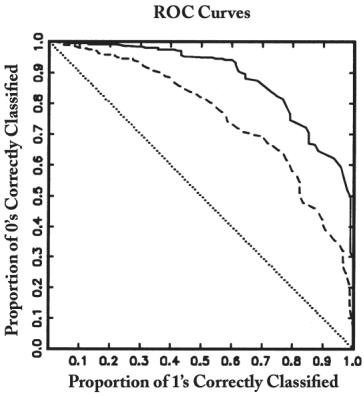
→ All these things are related, so some authors prefer different pairs of performance quantities [sigh]

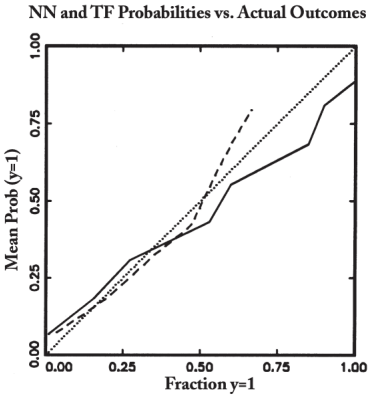
Traditionally, ROC curves plot recall and 1-precision

ROC


True positive rate
False positive rate
0
0.2
0.4
0.6
0.8
1
NetChop C-term 3.0
TAP + ProteaSMM-i
ProteaSMM-i

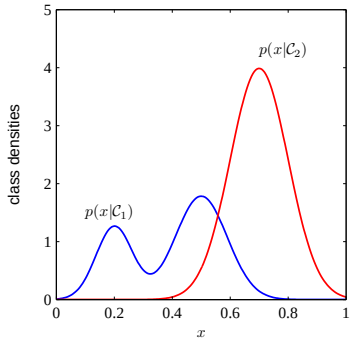
ROC AND CALIBRATION

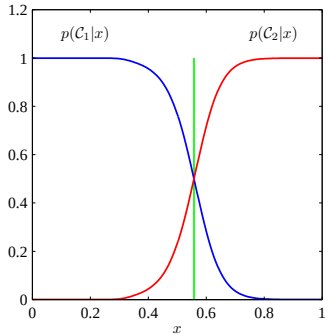




Decision boundaries

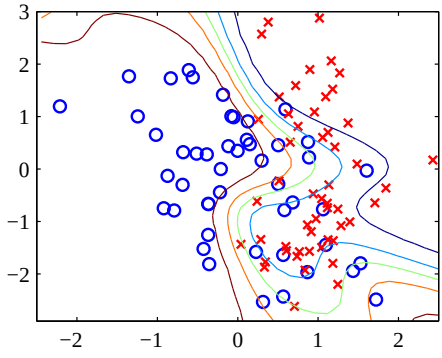
However we decide to set this threshold a classification model partitions $X_1 \ldots X_K$ into regions, based on what it would assign Y





Simple models generate simple decision boundaries

Decision boundaries



More complex models generate more complex decision boundaries

→ and need regularizing more carefully

Decision boundaries

The bias of a classifier determines the shape of the boundaries it can make

- → Linear models, e.g. additive logistic regression can make straight dividing lines
- → Neural networks make smooth curves

Both focus on learning a function

Decision boundaries

The bias of a classifier determines the shape of the boundaries it can make

- → Linear models, e.g. additive logistic regression can make straight dividing lines
- → Neural networks make smooth curves

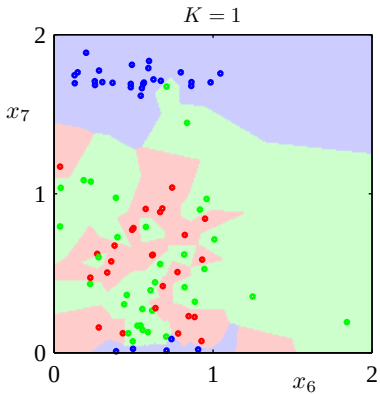
Both focus on learning a function Alternatively we can *ask the data*, e.g. the k-nearest neighbour classifier (for 0/1 classificatin):

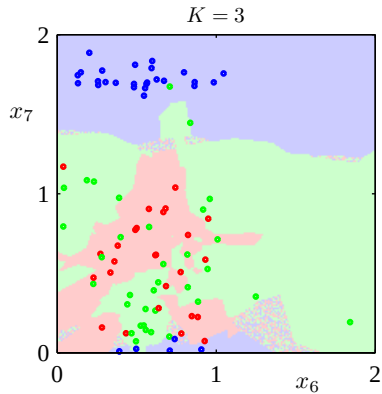
- → Takes your new data point
- → Finds the k nearest training observations it has seen
- → Asks each training observation for its class
- → Returns the proportion of those cases that were $Y = 1$

This is not great

- → but weirdly never more than twice as bad as the best possible classifier

Nearest neighbours





Adaptive neighbourhoods

If we want more control over k can use the axes to define the regions over which averaging happens

→ Split an axis (branch the tree) when the Y-averages on either side are *too different*

Last week we say that this generates quite *high variance* trees, so control overfitting by

→ resampling the data
→ fitting a new tree
→ averaging all the trees' predictions (the forest)

This (at a high level) is the 'random forest' model we saw last time

So what makes the *causal forest* model causal?

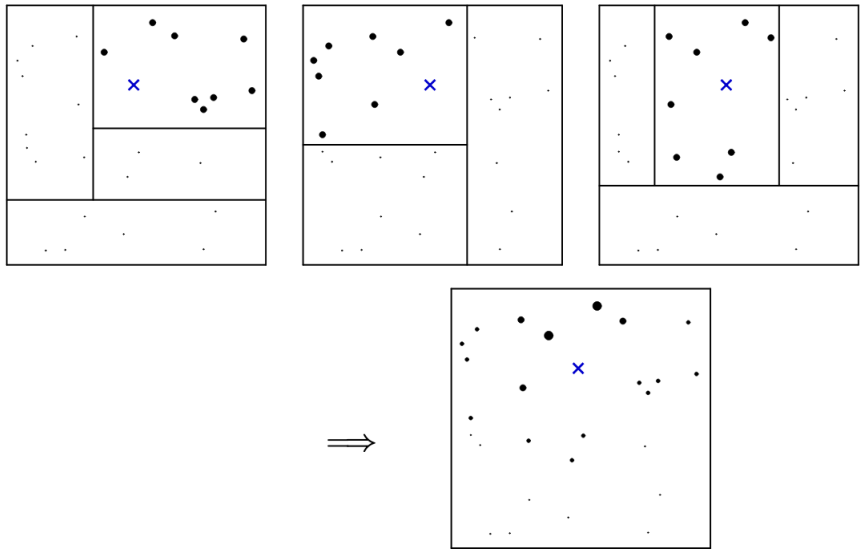
Causal forest

With nearest neightbour classification you are either averaging over some training case, or not

Athey et al.prefer to average over more, but weight them by distance (Athey et al., 2018)

→ the weight for a training case is the proportion of times it ends up in the same leaf of a tree (remember there are lots of trees in a forest)

Causal forest



Causal forest

Add

- → Separate models of the treatment variable and the outcome, then double-ML style fitting
- → Cross-fit to avoid overfitting
- → built-in propensity score weighting

Interestingly out of sample fit is no longer the criterion - it's causal inference

- → when you think about it, causal inference problems are always about out of sample (and sometimes out of world)

References

Athey, S., Tibshirani, J. & Wager, S. (2018). 'Generalized random forests'.

Broockman, D. E. (2016). 'Approaches to studying policy representation'. *Legislative Studies Quarterly*, *41*(1), 181–215.

Doyle, M. W. & Sambanis, N. (2000). 'International peacebuilding: A theoretical and quantitative analysis'. *American Political Science Review*, *94*(4), 779–801.

King, G. & Zeng, L. (2001). 'Improving forecasts of state failure'. *World Politics*, *53*(4), 623–658.

King, G. & Zeng, L. (2007). 'When can history be our guide? the pitfalls of counterfactual inference'. *International Studies Quarterly*, *51*(1), 183–210.