ADVICE FOR APPLYING MACHINE LEARNING

Debugging Machine Learning.

Mostly for profit but with a bit of fun too!

Michał Łopuszyński

PyData Warsaw, 19.10.2017

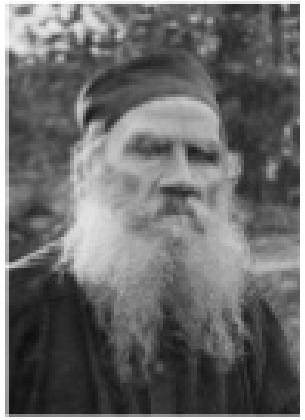

Hey, my
ML system
does not
work at all...

Hint #1
Check your code

AKA it is engineering, stupid!

Hint #2
Check your data

Data quality audits are difficult



Happy families are all alike; every unhappy family is unhappy in its own way.

Leo Tolstoy

Like families, tidy datasets are all alike but every messy dataset is messy in its own way.

Hadley Wickham

H. Wickham, Tidy Data, JSS 59 (2014), doi: 10.18637/jss.v059.i10

Images credit - Wikipedia

Data quality

- Beware, your data providers usually overestimate the data quality
- Understand your data
  - Do exploratory data analysis
  - Visualize, visualize, visualize
  - Talk to the domain expert
- Think of outliers, missing values (and how the are represented)
- Is your data correct, complete, coherent, stationary (seasonality!), deduplicated, up-to-date, representative (unbiased)


OK, my ML system works, but I think it should perform better...

Hint #3
Examine your features

Features

- Features make a difference!
- Understand what features are important for your model
  - Use ML models offering feature ranking
  - Use feature selection methods
- Be creative with your features
  - Try meaningful transformations, combinations (products/ratios), decorrelation...
  - Think of good representations for non-tabular data (text, time-series)
  - Make conscious decision about missing values

Hint #4
Examine your data points

Data points

- Find difficult data points! (DDP)
- DDP = notoriously misclassified (or high error) cases in your cross-validation loop for large variety of models
- Examine DDPs, understand them!
- In the easiest case, remove DDPs from the dataset (think outliers, mislabeled examples)

Data points

- Get more data!
- Good performance booster, rarely applicable
- Trick 1. Extend your set with artificial data
  E.g., data augmentation in image processing,
  SMOTE algorithm for imbalanced datasets
- Trick 2. Generate automatically noisy labeled
  data set by heuristics, e.g. distant supervision in NLP
  (requires unlabeled data!)
- Trick 3. Semisupervised learning methods
  self-training and co-training (requires unlabeled data!)

Hint #5
Examine your model

Why my model predicts what it predicts? (philosophical slide)

- **How do** you answer **why** questions?
- Inspiring homework: watch Richard Feynman, Fun to imagine on magnets (youtube)

Model introspection

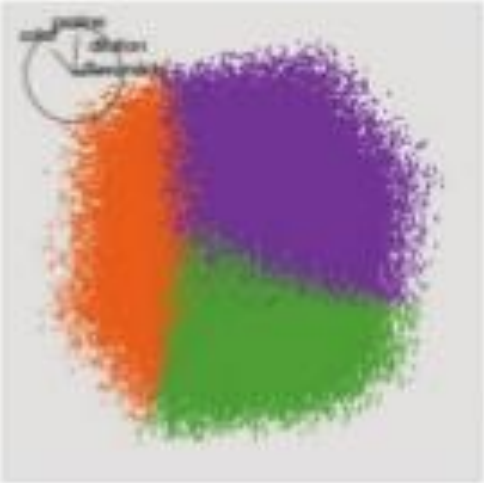
- You can answer thy why question, only for very simple models (e.g., linear model, basic decision trees)
- Sometimes, it is instructive to run such a simple model on your dataset, even though it does not provide top-level performance
- You can boost your simple model by feeding it with more advanced (non-linearly transformed) features

Visualizing models

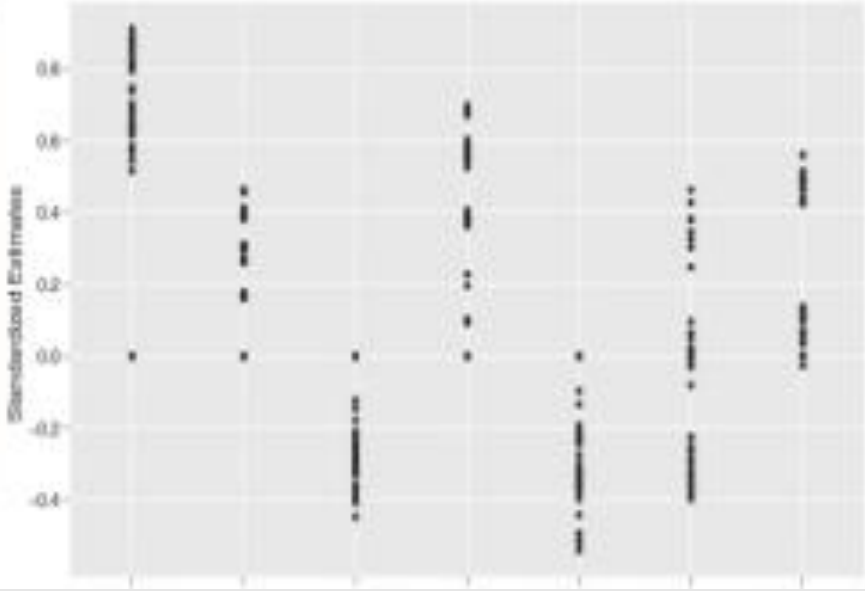
Visualizing Statistical Models: Removing the Blindfold

Hadley Wickham,[1*] Dianne Cook,[2] and Heike Hofmann[2]

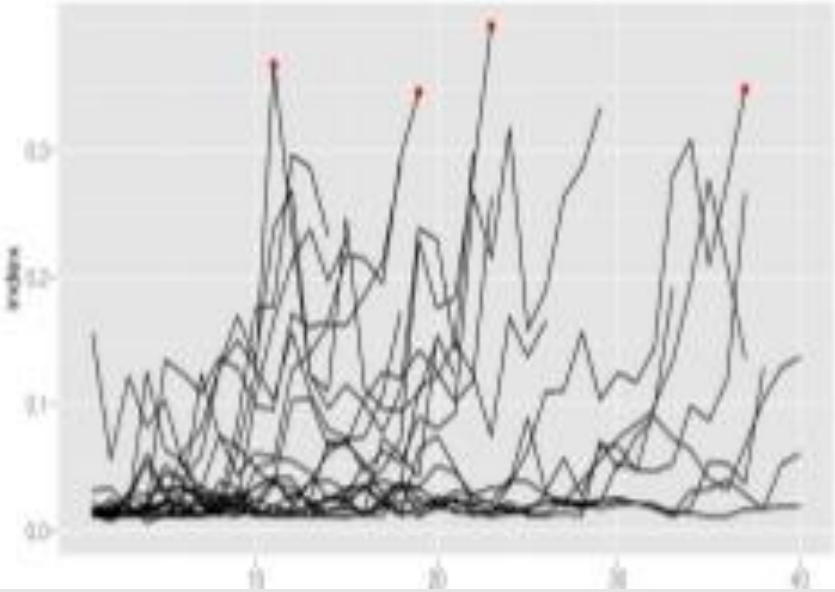
Display model in a data space



Look at collection of models at once



Explore the process of model fitting




So my ML system works on test data, but you tell me it fails in production?

Hint #6
Watch out for overfitting

Overfitting

If you torture the data long enough, it will confess.

Roald Coase

Hint #7
Watch out for data leakage

Data leakage

- Some time ago, I used to thing data leaks are trivial to avoid
- They are not! (Look at number of Kaggle competitions flawed by Data Leakage)
- You may introduce them yourself
  E.g. meaningful identifiers, past & future separation in time series
- You may receive them in the data from your provider
- Good paper

Leakage in Data Mining:
Formulation, Detection, and Avoidance

Shachar Kaufman Saharon Rosset Claudia Perlich

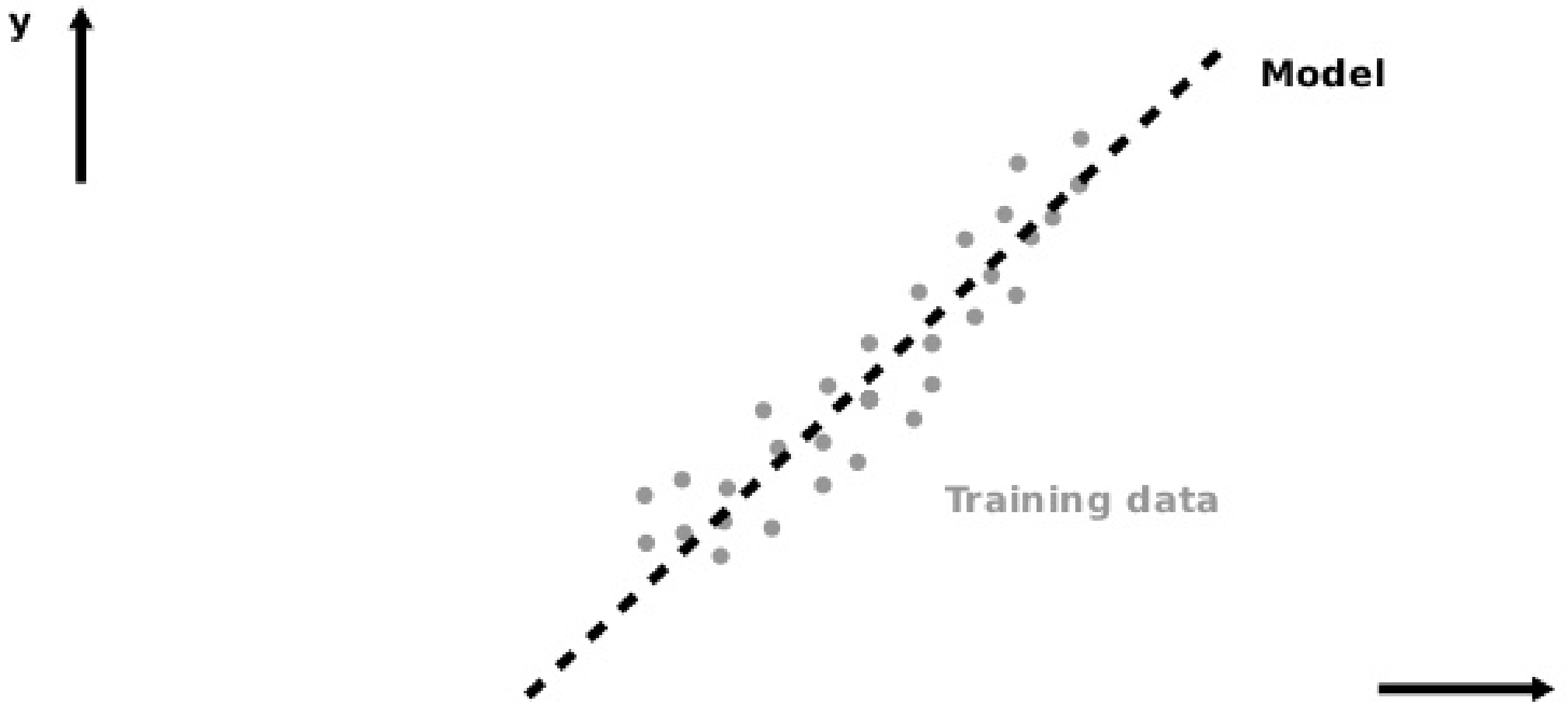
Example

- Label as feature
- Feature with direct relation to Label
- New data added with time

Hint #8
Watch out for covariate shift

What is covariate shift?

What is covariate shift?



What is covariate shift?

What is covariate shift?

Covariate shift

- Unlike overfitting and data leakage, it is easier to detect
- Method: Try to build classifier differentiating train from production (test).
  If you succeed, you very likely have a problem
- Basic remedy – reweighting data points. Give production-like data higher impact on your model

…

- Correct Sampling
- Random Testing


The quality of my super ML system deteriorates with time, really?
Really really?

Hint #9
Remember monitoring & maintenance

AKA it is engineering <u>again</u>, stupid!

Source

- https://www.slideshare.net/lopusz/debugging-machinelearning