Practical AI: ML as a framework

Stanislav Protasov for
Harbour.Space University

Agenda

Problems suitable for ML

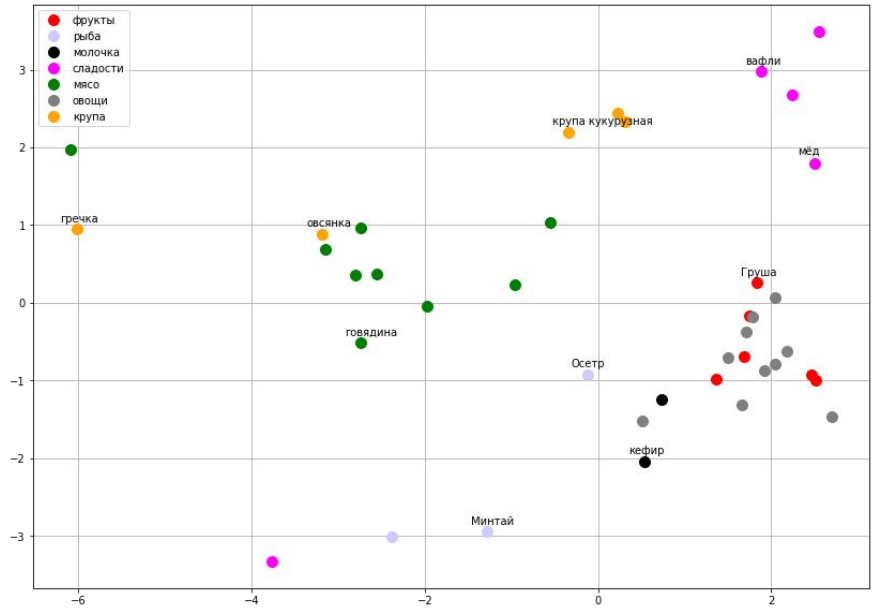
Steps of ML solution

False friends of ML

Sequential data

Is there a problem for ML?

Visualization of 4+D data



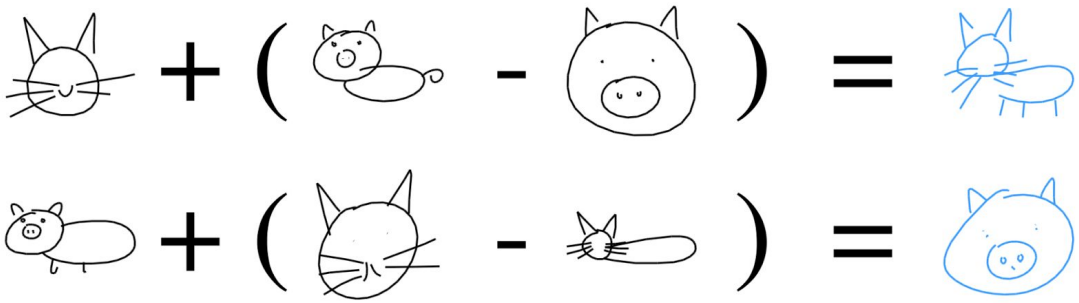
Rule of thumb:
clustered data should remain clustered

PCA (principal component analysis, with SVD)

LDA (latent Dirichlet allocation) — considers document (sample) as a set of “related to” topics

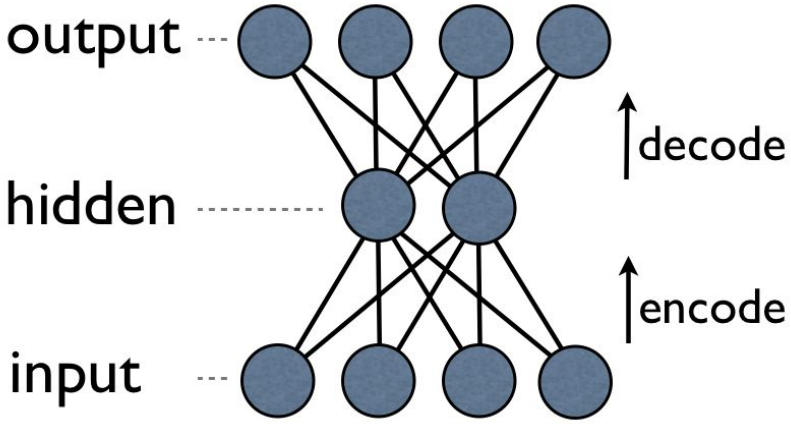
t-SNE - best for visualization

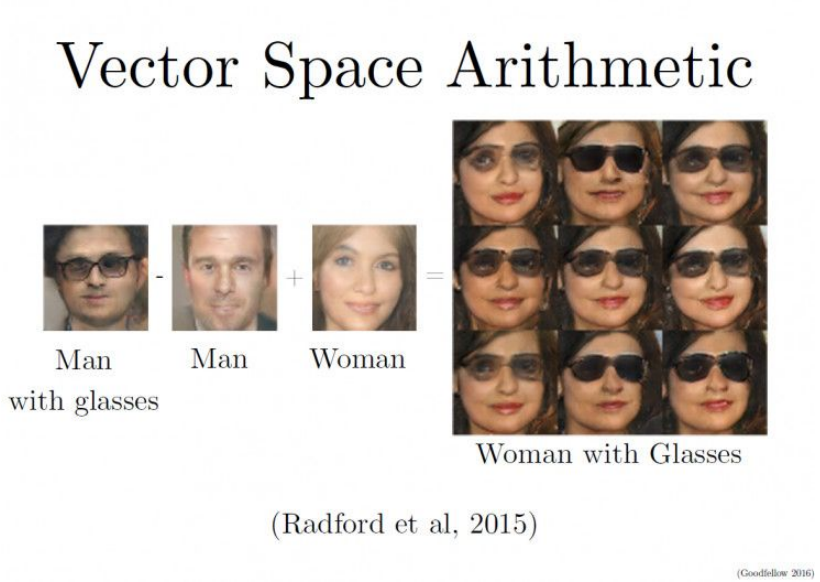
Embedding



PCA

Autoencoders



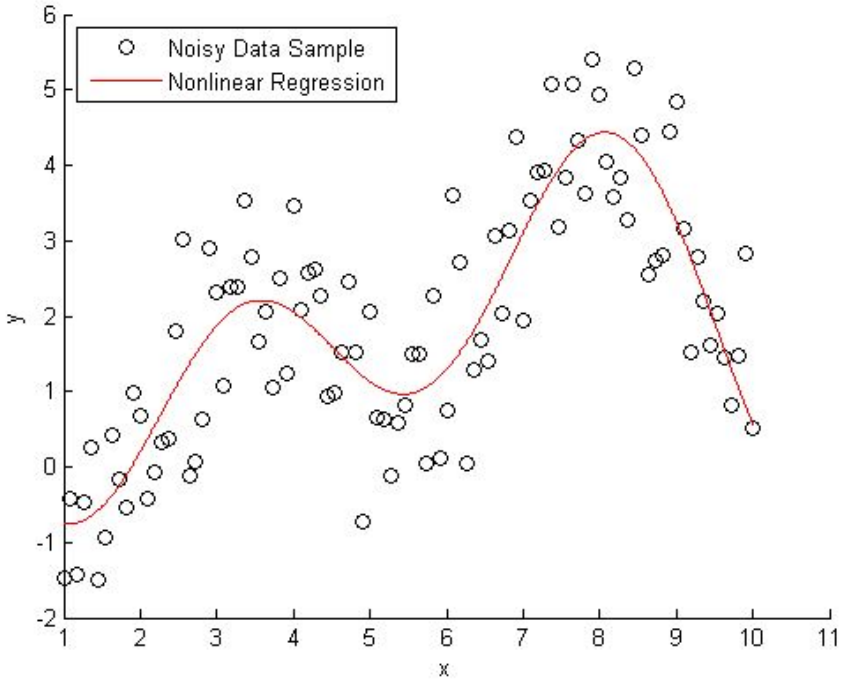


See also http://www.offconvex.org/2016/02/14/word-embeddings-2/

Prediction of values and probabilities

Regression can be considered as **scoring** the data (prediction of values)

1) [Linear] Regression
    a) With GD
    b) With LSA
    c) With ...
2) SVM (with kernels)
3) HMM
4) ANNs



Separating data into **groups (tagging)**

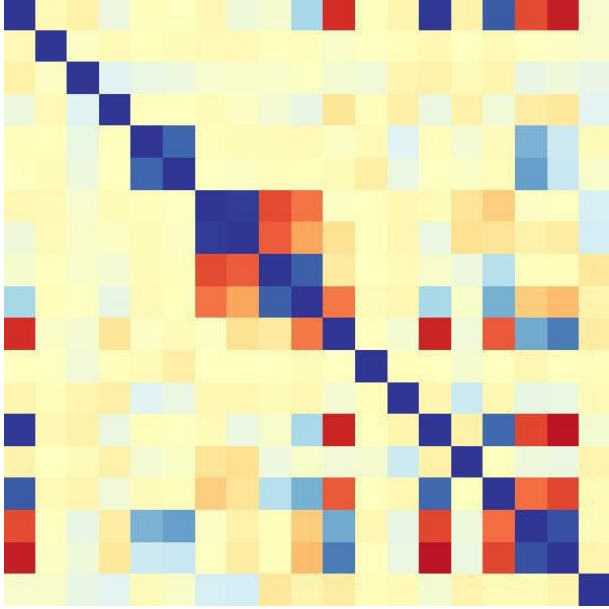
1) ***K***-class **classification** is usually a function

$$F : \Omega \to [0..1]^K$$

2) ***K***-cluster **clustering** of ***N*** objects is usually a function

$$F : \Omega^N \to \{1, \ldots, K\}^N$$

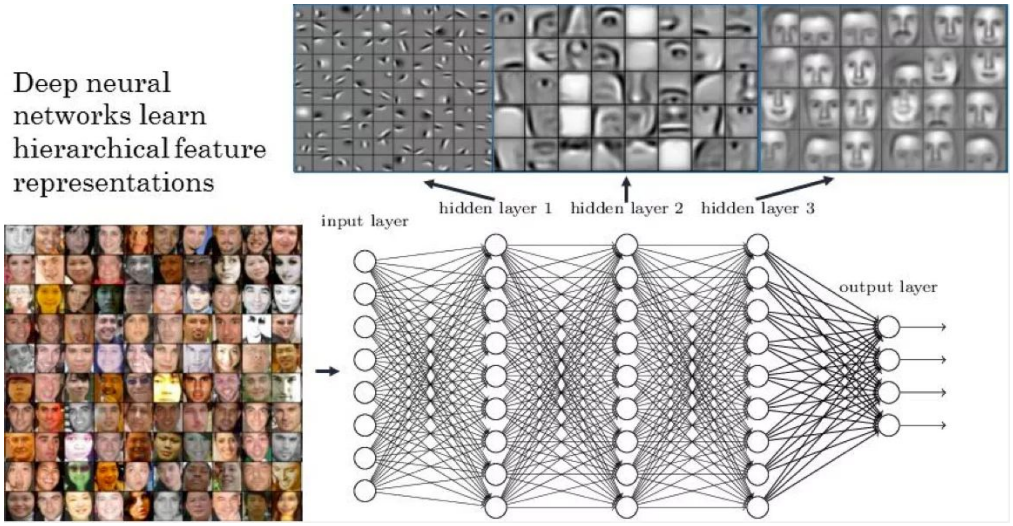
Explaining the data and model



1) **Factor analysis** with **covariance matrix** is a good way of analyzing **factors** (features).

2) **Linear and tree models** are highly explainable (linear and logistic regression, LSA, ...).

3) **ANNs can be explained** much harder, but still can be.



Important steps towards solution

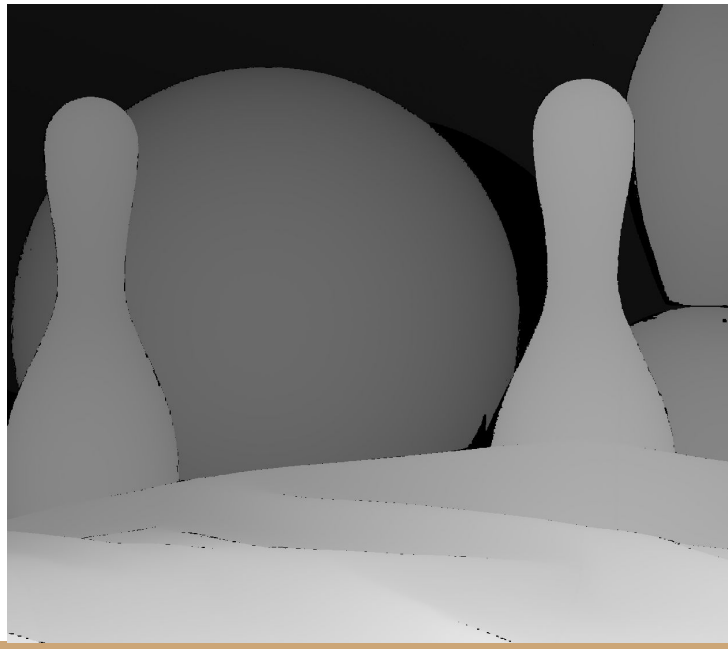
Top-level overview of how ML models are created

1. Find data
2. Prepare data
3. Prepare dataset
4. Train, validate, test
5. Measure quality
6. Save, deploy
7. Improve

1. Find your dataset

Dataset = samples + target (or ground truth)

1) Collect data for your task
2) Take the data from customer
3) Download publicly available dataset



1.1. Dataset and quality

Before you start training the model, be sure you understand:

- How do you measure the quality?
- CAN YOU?
- What are the values that will satisfy you?

Lab #1: Explaining the model, measuring quality

1) Explore naive-ml example.
    a) Consider difference between matrix inverse and LSA.
    b) Compute RMSE for both solutions
        i) Which of solutions is more accurate?
2) Find an approximation for GPD. Compute RMSE

2.1. Clean your data

- Clean
- Restore nulls
- Normalize
- Extend
- Augment
- Bootstrap
- ....

2. Split your data for training

1. **Train**
2. **Validate**
3. Test

Firstly your model is trained to **minimize error** on **training set**.

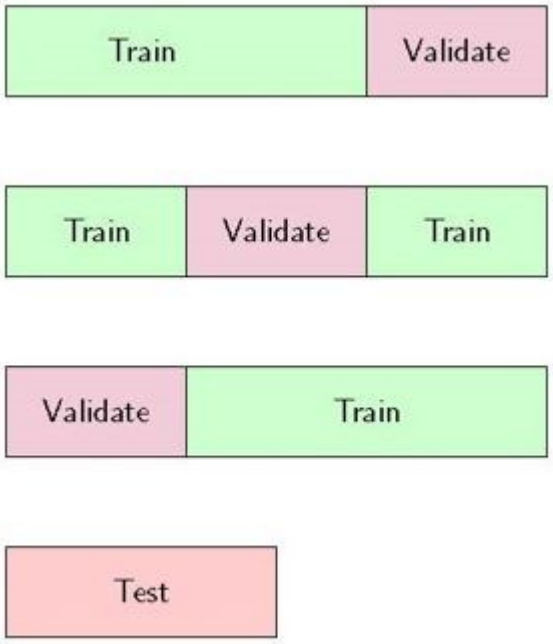
Validation data is used to (1) prevent overfitting (2) tune hyperparameters.

Parameters and hyperparameters that minimize error for **validation set** are desired result.

Test set is used to compute **quality results**. (Consider this as blind **acceptance** by customer).

... or

1) Split you data into train+validate and test sets.
2) Use **cross-validation** for tuning **parameters**
3) Use grid/random/... **search** for tuning **hyperparameters**.



3. Train your model and save results

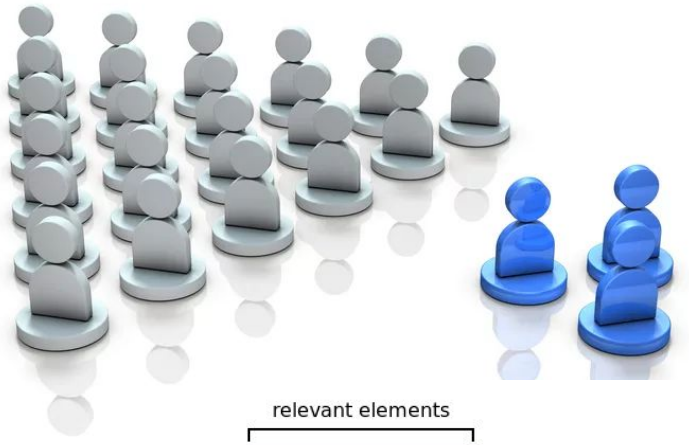
The results of your training (the most valuable thing!):

- Model type (ANN, SVM, CNN, R-CNN, ...)
- Hyperparameters
- Parameters (weights)

SAVE THEM IF YOU LIKE THEM

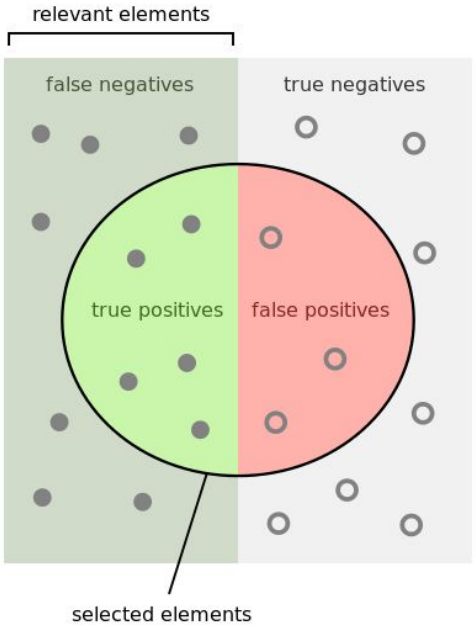
False fiends of ML

Biased data



WRONG:

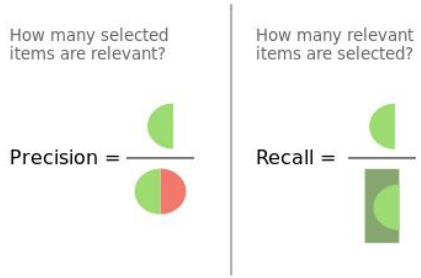
Quality = Accuracy = (TP+TN)/(P+N)



BETTER:

Precision, Recall

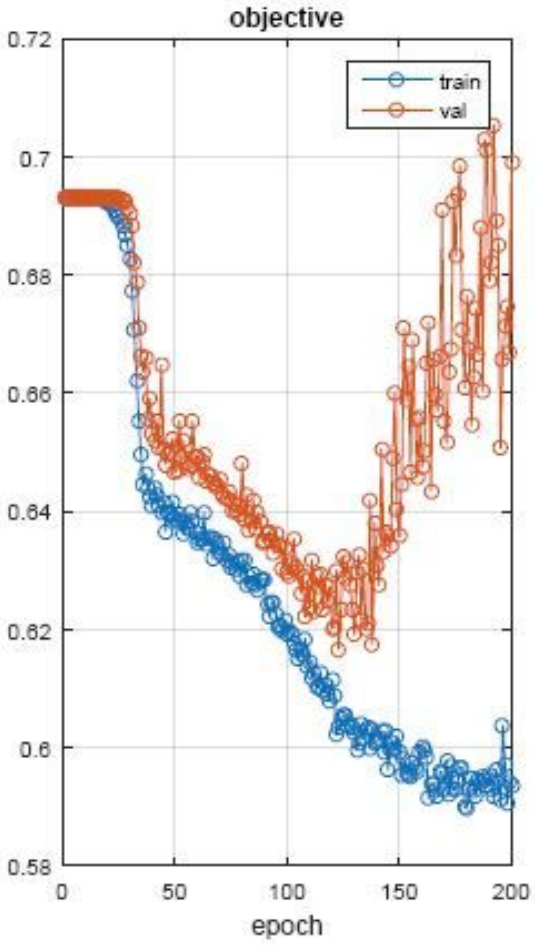
Quality = $F_1 = \frac{2}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$

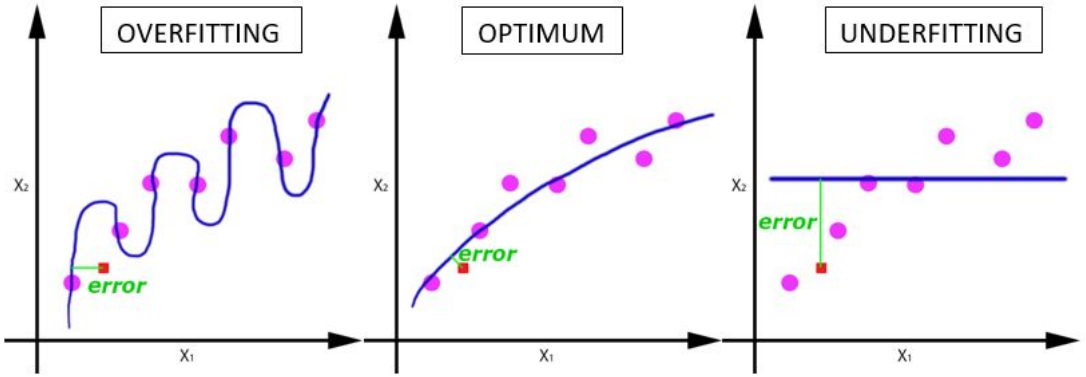


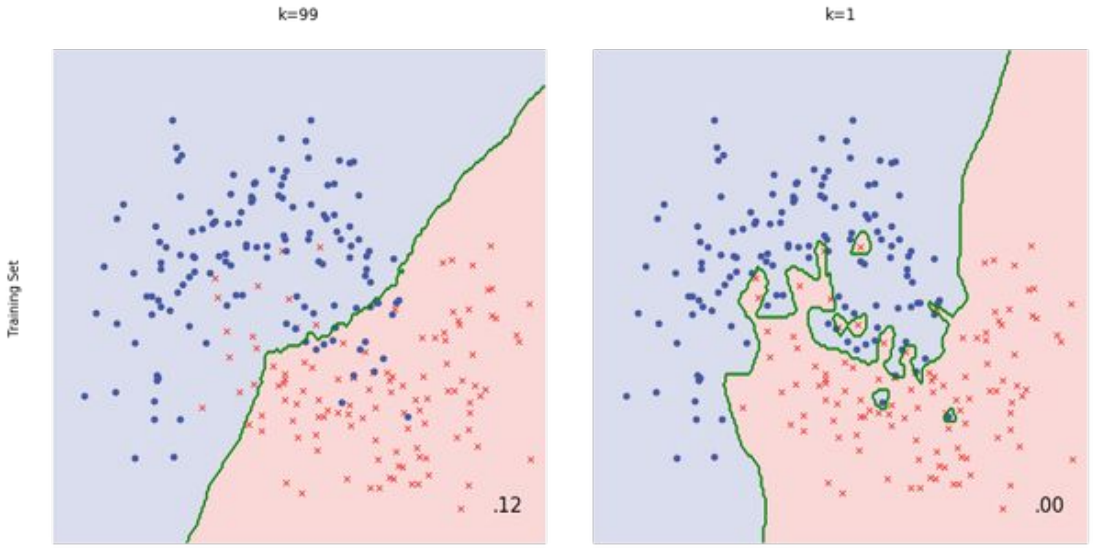
EVEN BETTER:

Normalize your data distribution
(find examples, augment, or at least clone)

Overfitting







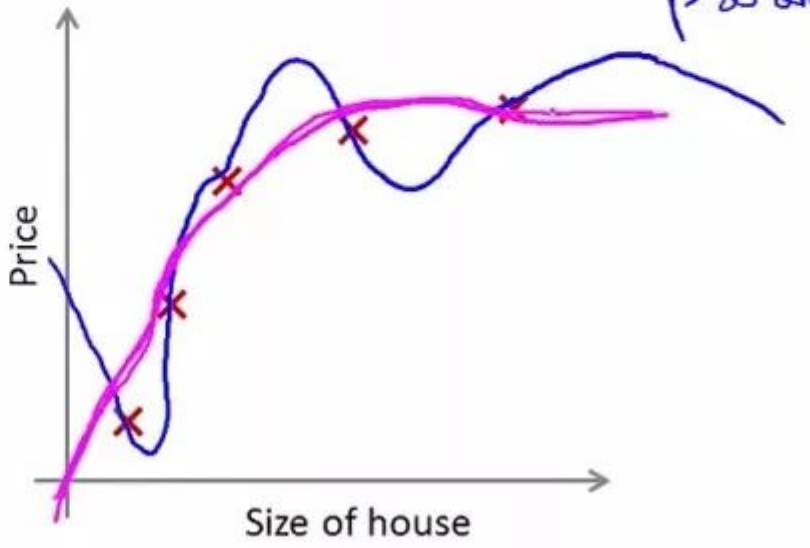
Overfitting?

Regularization.

$$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^{n}\theta_j^2\right]$$

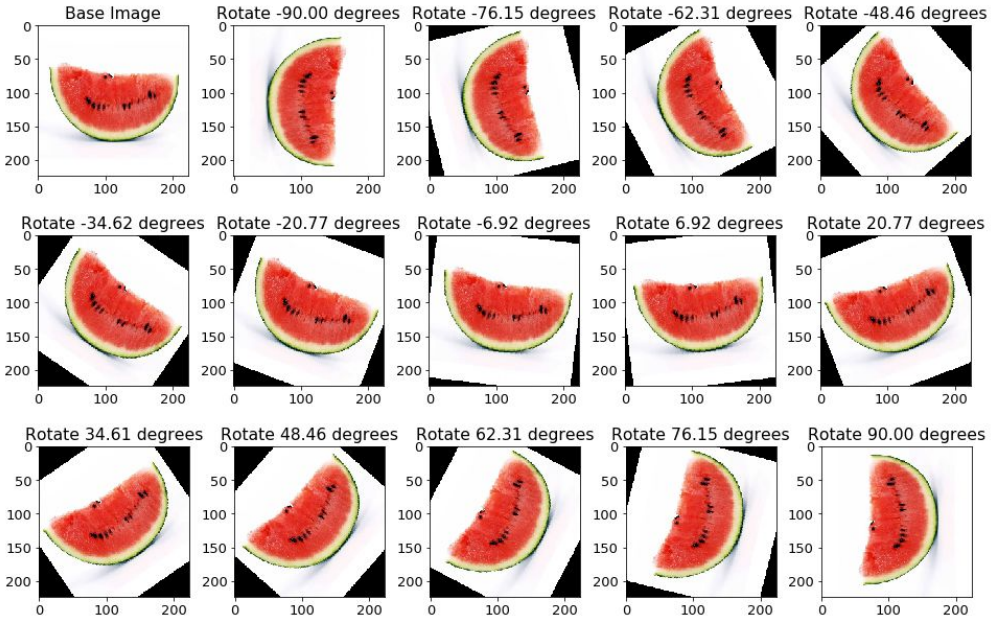
regularization parameter

$$\min_\theta J(\theta)$$

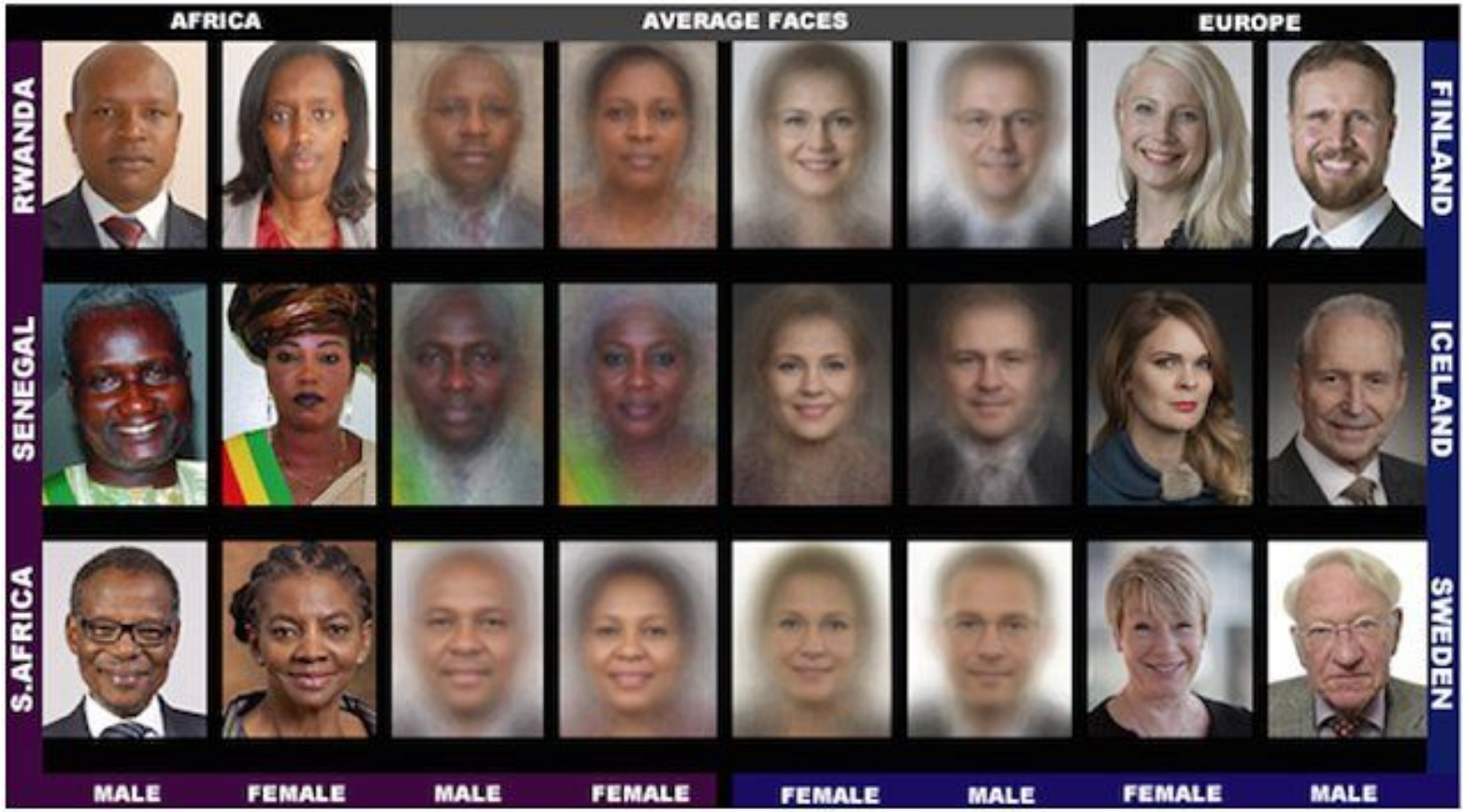


Small dataset and complex model

1) Don’t use complex model for small dataset
   a) Rule of thumb: number of **parameters** should be comparable with **dataset size**
2) Data **augmentation**
3) Data generation



Biased conditions (datasets)



ML for CV

Machine learning is...

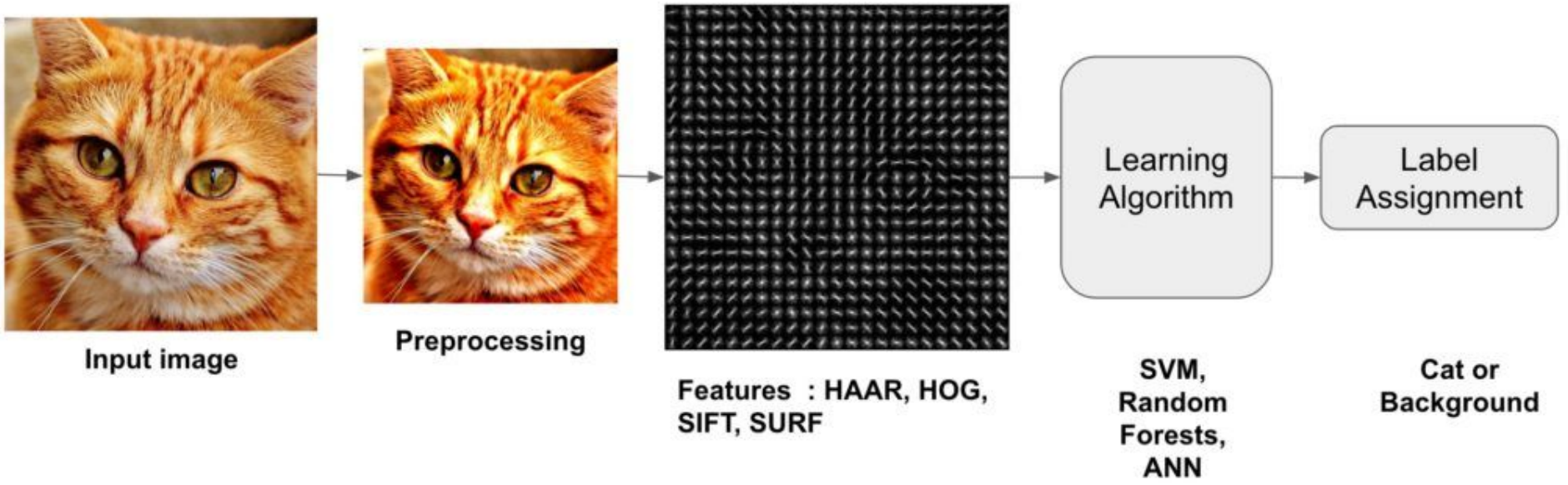
Finding a **function** over some sample **space** by **examples**

Function: classifier, regression (dimension reduction)

Space: image itself (for deep learning), feature space for classical ML

Examples: multiple examples of images of desired objects

Classifier example



Graded lab #10

Select one of problems and submit to Canvas:

1. How many red and yellow stones are on this image? https://github.com/hsu-ai-course/hsu.ai/tree/master/code/datasets/images/curling.jpg
2. Use this https://github.com/jhlau/doc2vec pretrained model to build embeddings of texts:
   a. `This sentence is about fish and sea`
   b. `How much should I pay for this fish? Sounds like it just was caught in the sea!`
   c. `Integration is opposite to derivation.`
   d. `Sine function derivative is cosine function.`

   What are pairwise cosine similarity values?

Homework

Start reading this book - this is a cool starter for ML.

Demonstrate following skill (mandatory, edvanced and facial=nightmare modes):

- splitting data to test and validate sets,
- introducing and measuring error (cost) function for your data,
- cross-validation.