Primer on ML [1]

Overview

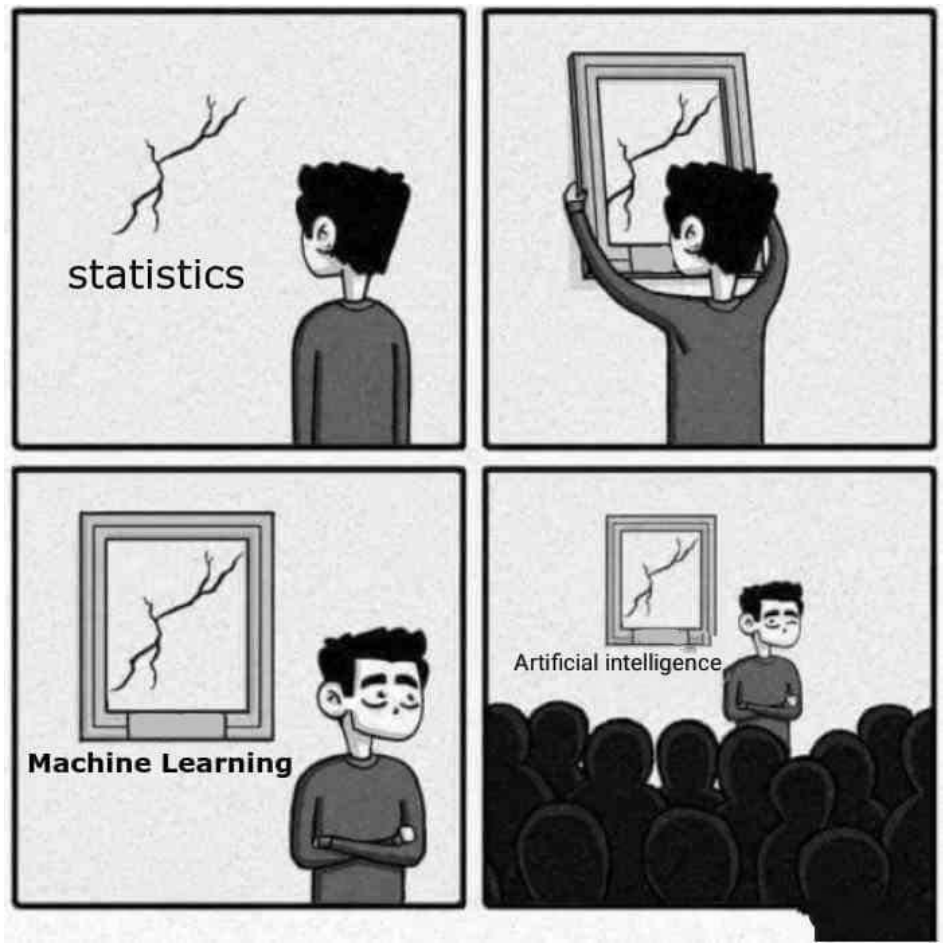
- What is ML
  - supervised [s]
  - unsupervised [u]
  - reinforcement learning [rl]
- Gently into [s]
  - linear models
  - trees
- ML specifics
  - metrics
  - overfitting
  - regularization
  - model validation

based on and inspired by

- MIPT x Yandex course [1]
- ML in HEP summer school [2]
- Data Mining in Action [3]

General concepts

What is ML



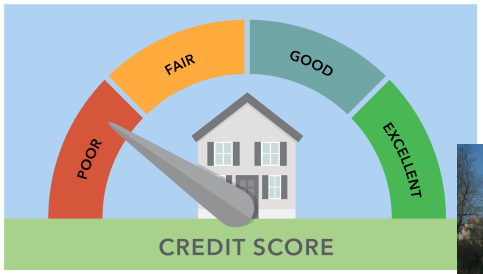
What is ML

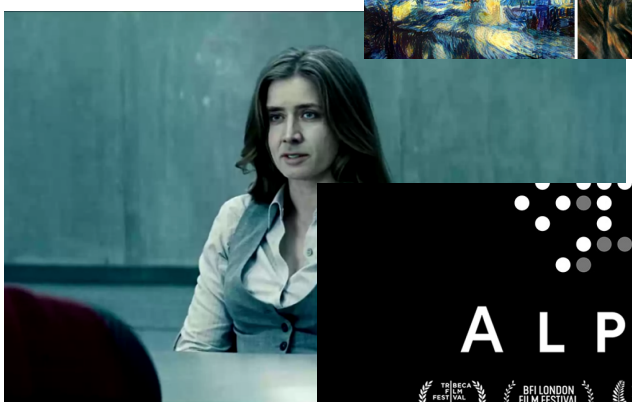
Words

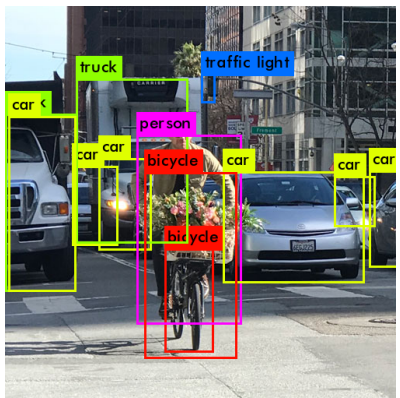
- Science of getting computers to act without being explicitly programmed [Coursera]
- Scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead [Wikipedia]
- Science searching for hidden relations in data [A. Volokhova]

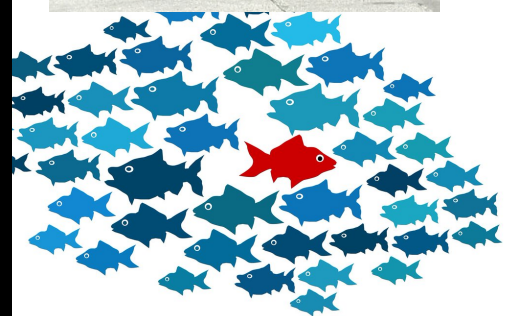
What is ML

Pics

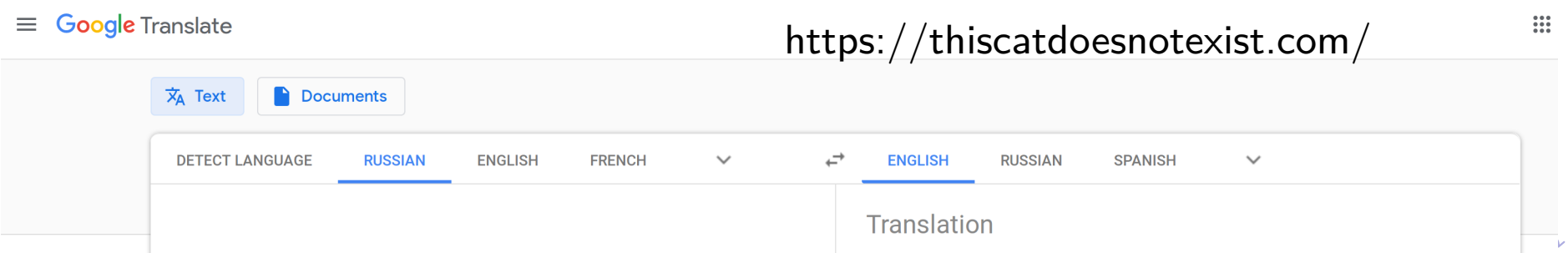






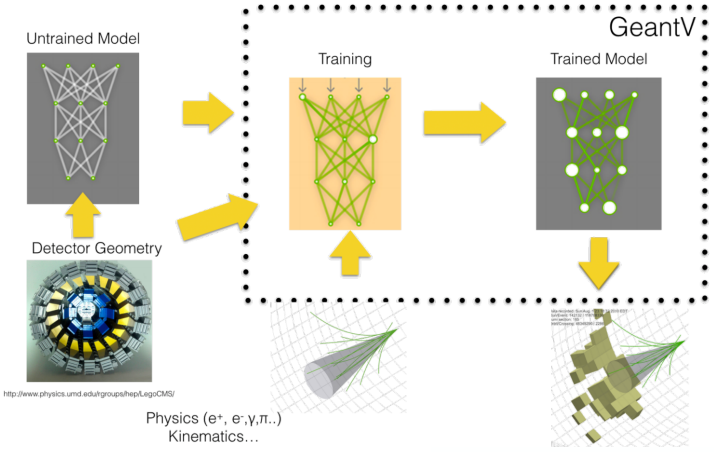


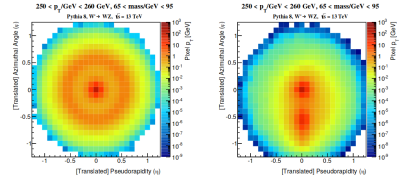
https://thiscatdoesnotexist.com/

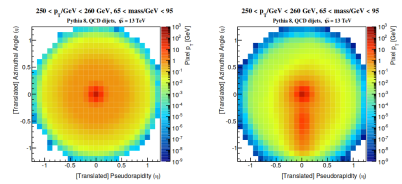


What is ML

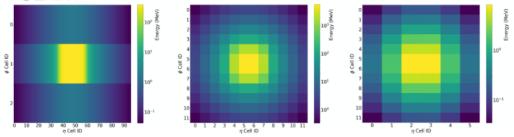
HEP pics

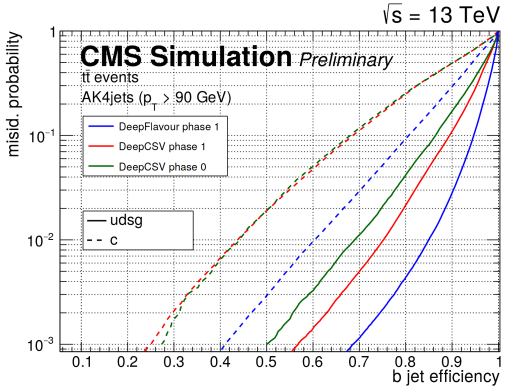


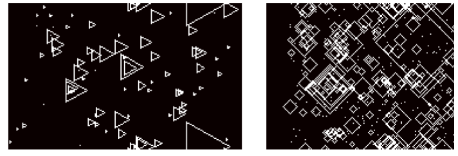




- GEANT

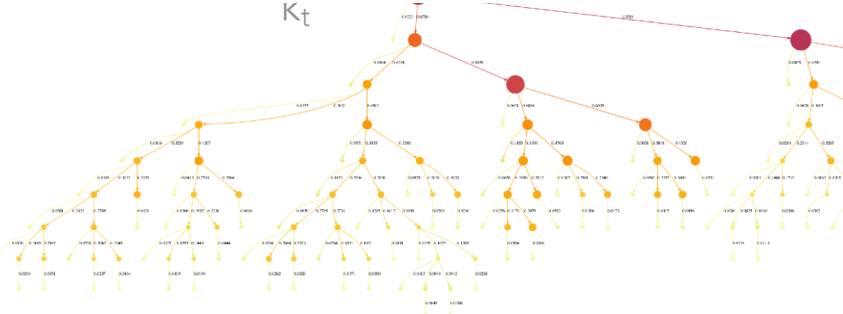


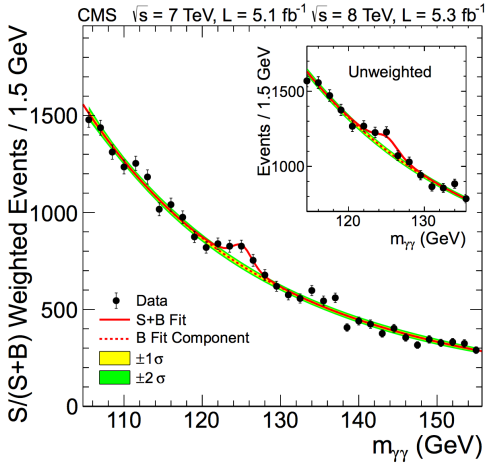


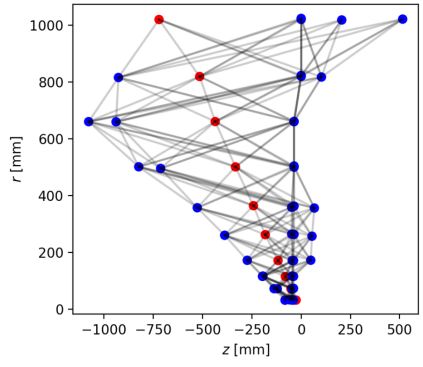


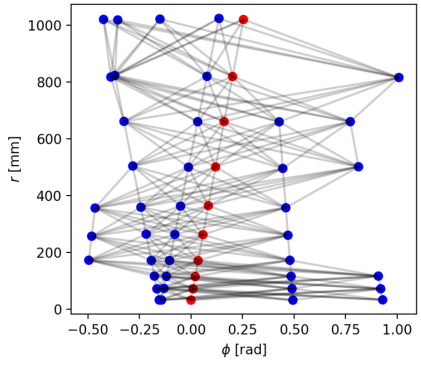
(a) Photons (b) Charged Particles

(d) Lepton









ML branches

- Supervised
  - Classification
  - Regression
- Unsupervised
  - Clustering
  - Anomaly detection
  - Generative models
  - Dimensionality reduction
  - ...
- Reinforcement Learning

Supervised

Definitions

Feature space $\mathbb{X}$

- real
- categorical

Target space $\mathbb{Y}$

- $\mathbb{Y} = \overline{(0,C)} \rightarrow$ classification
- $\mathbb{Y} = \mathbb{R}^d \rightarrow$ regression

Sample $(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_\ell)$

Training set $X,Y = (\mathbf{x}_i, \mathbf{y}_i)_{i=1}^{\ell}$

Model $a(\mathbf{x}) \leftarrow$ predicts target

Loss function $L(a,Y) \leftarrow$ evaluates model's quality

Training

$$a(\mathbf{x}) = \operatorname{argmin}_{\mathbf{a} \in \mathbb{A}} L(a,Y)$$

Supervised

C vs R

? cat, dog or muffin

? relevant or spam

? disease or not

? good or bad

- rent price
- temperature
- annual profit
- driving time

Supervised

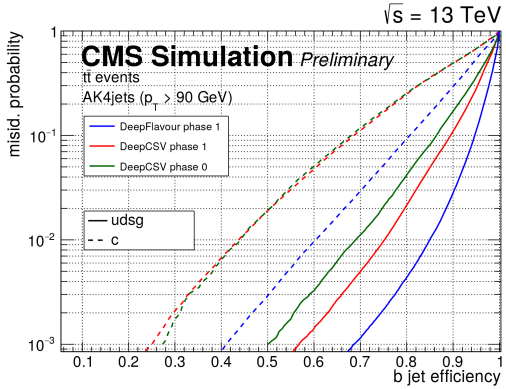
Classification in HEP

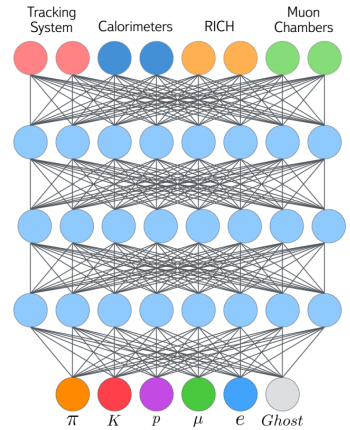
? b, c or ud jet

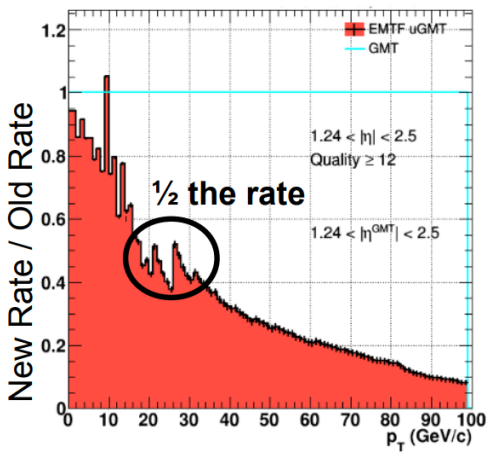
? π, K, or μ particle

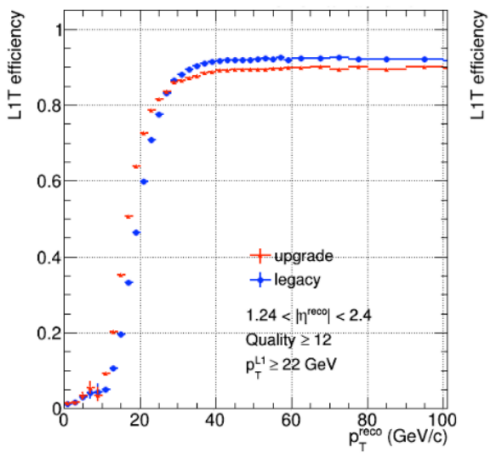
? $t\bar{t}$ or QCD event

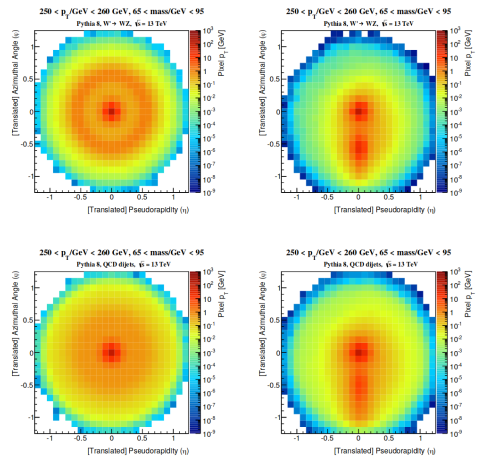
? select or reject trigger candidate







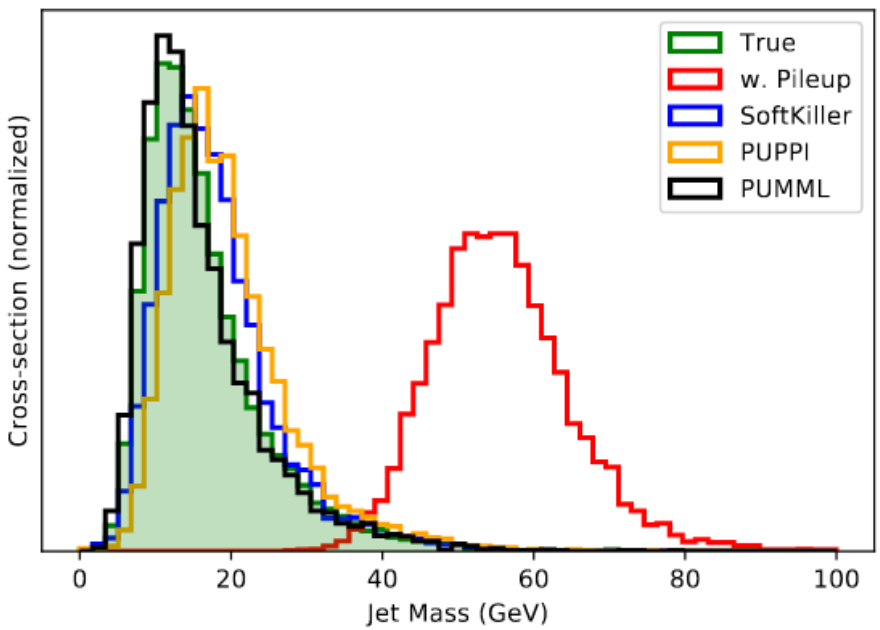


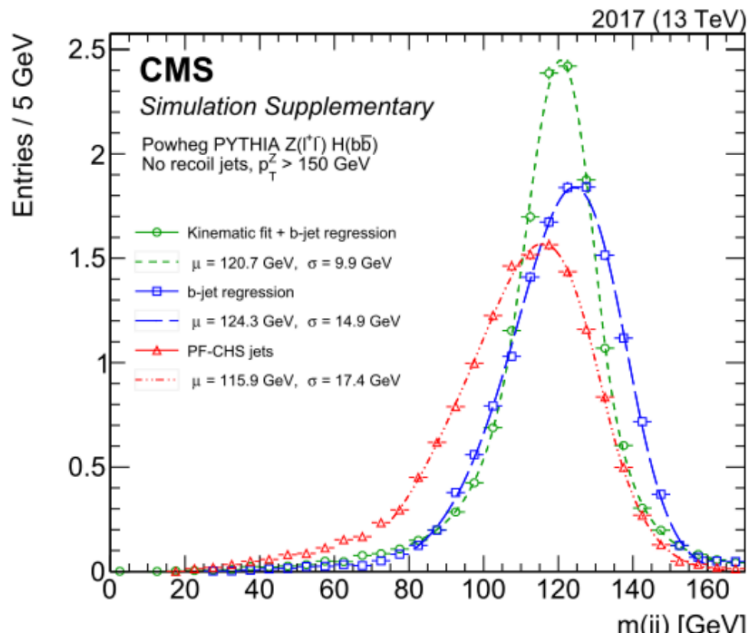


Supervised

Regression in HEP

- energy resolution [1]
- PU mitigation [2]



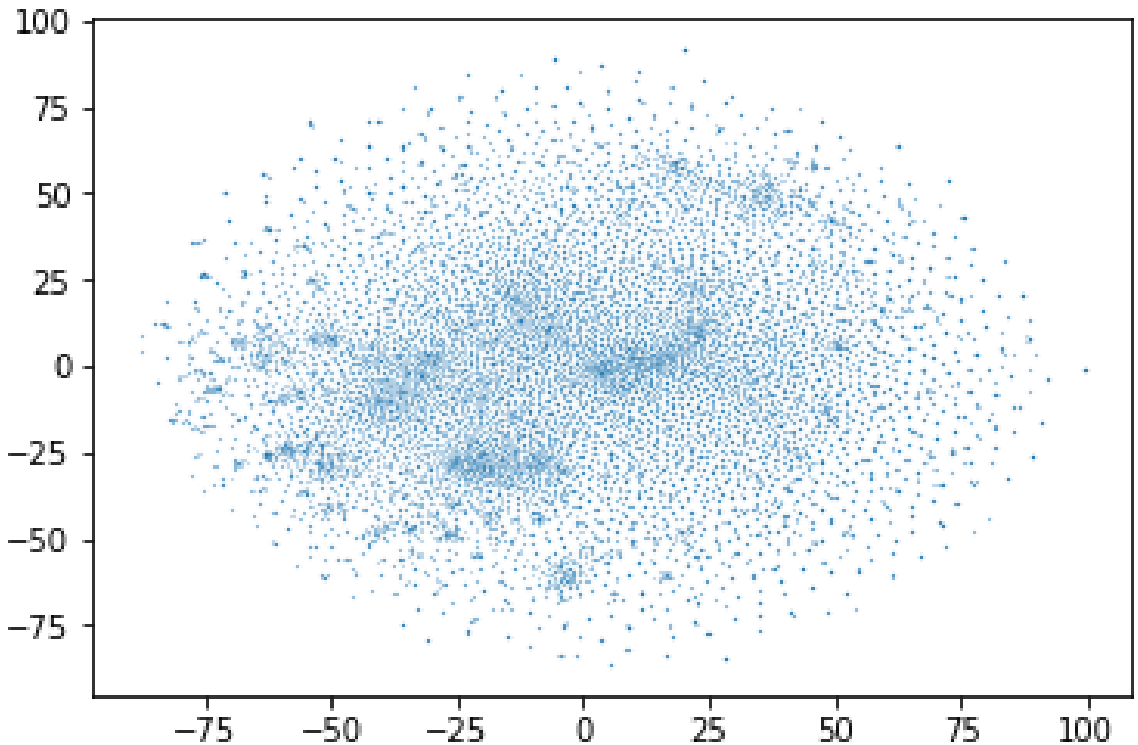


ML branches

- Supervised
  - Classification
  - Regression
- Unsupervised
  - Clustering
  - Anomaly detection
  - Generative models
  - Dimensionality reduction
  - ...
- Reinforcement Learning

Unsupervised

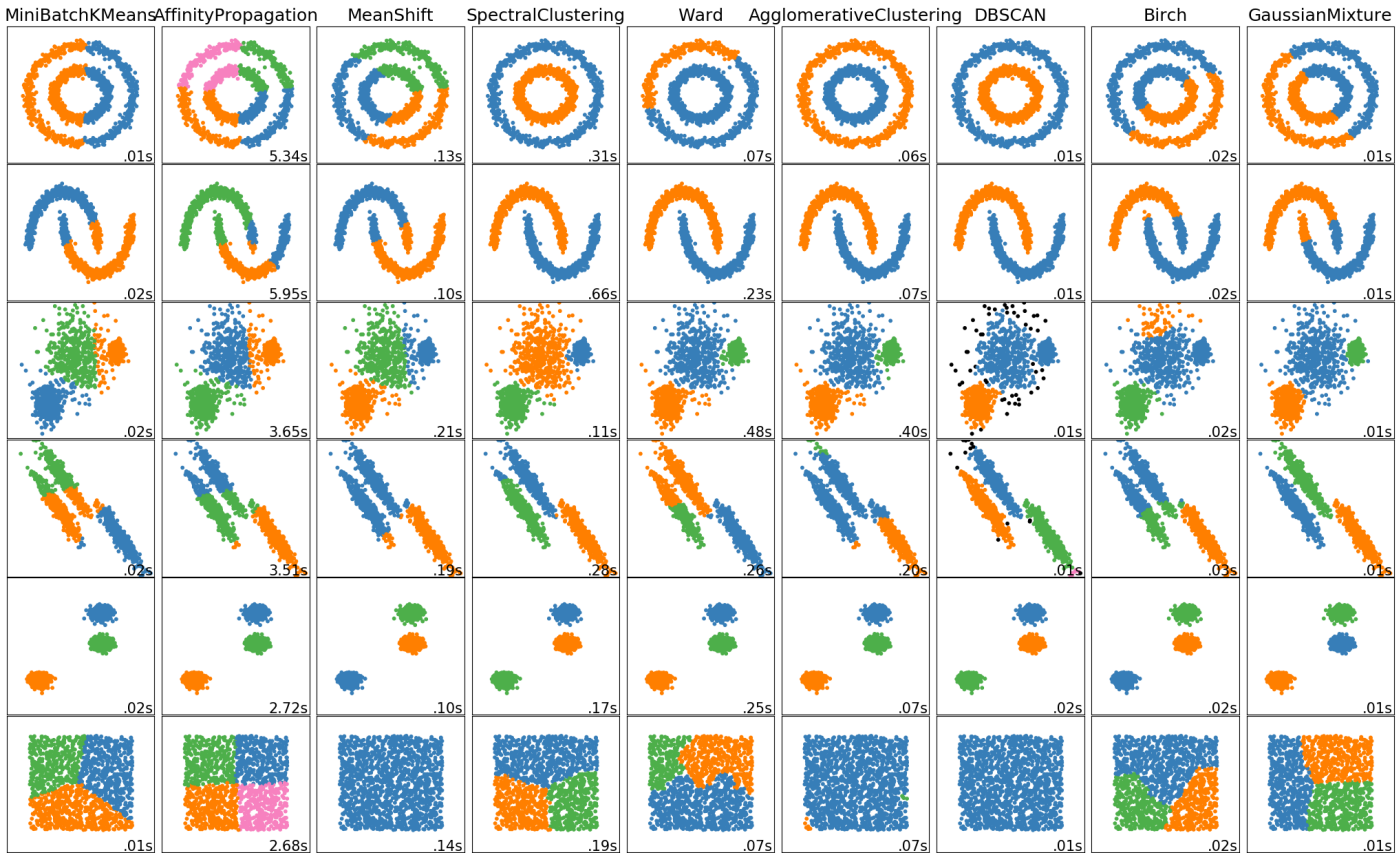
Clastering



visuals! [link]

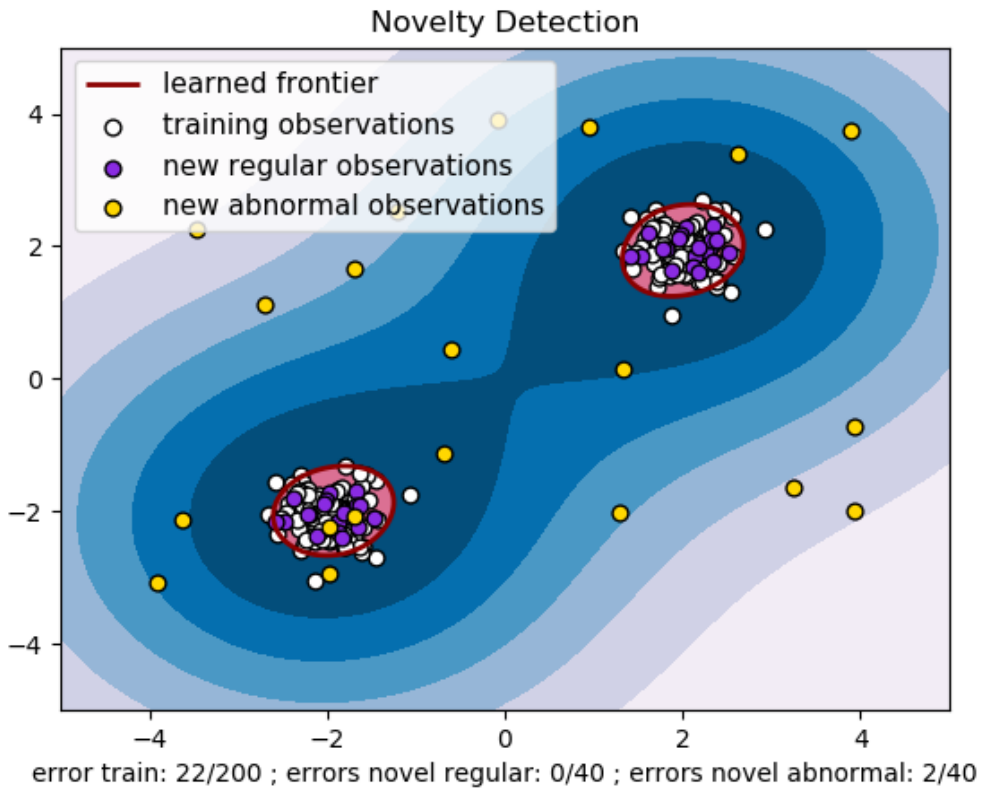
Unsupervised

Clastering



Unsupervised

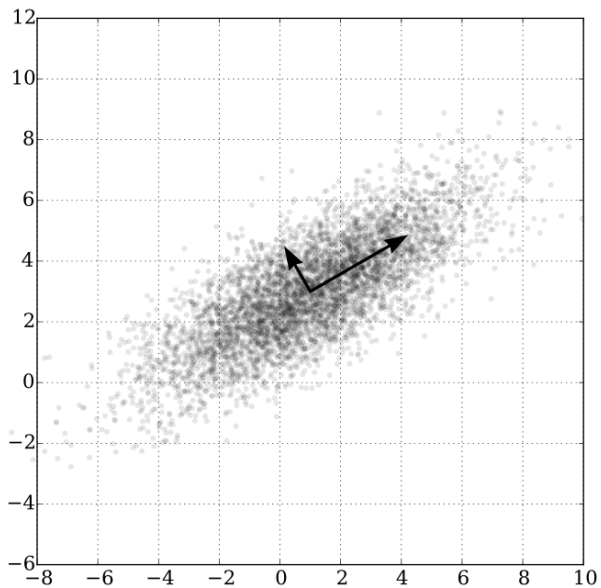
Novelty & Outliers

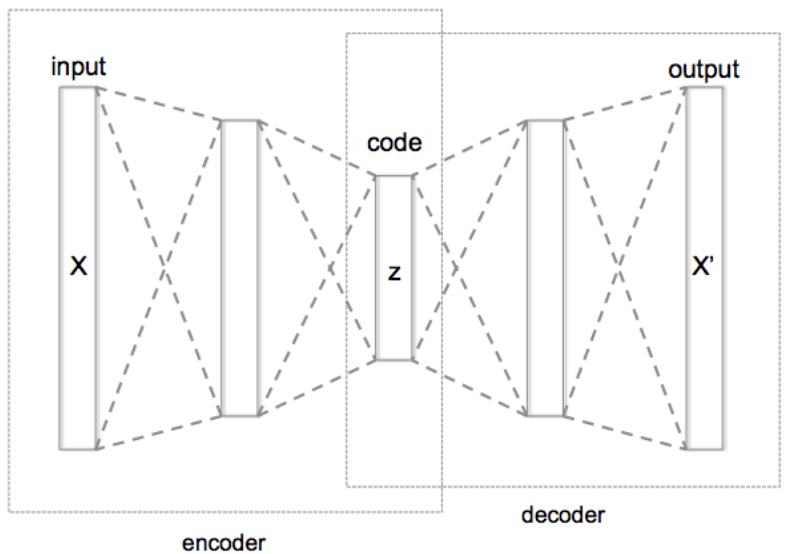


Unsupervised

- Dimensionality reduction
  - Principal component analysis
  - Autoencoders
- ∗ Generative models
  - GANs, VAE

more visuals! [link]

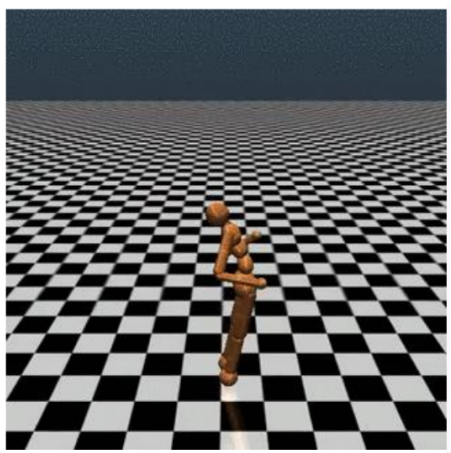


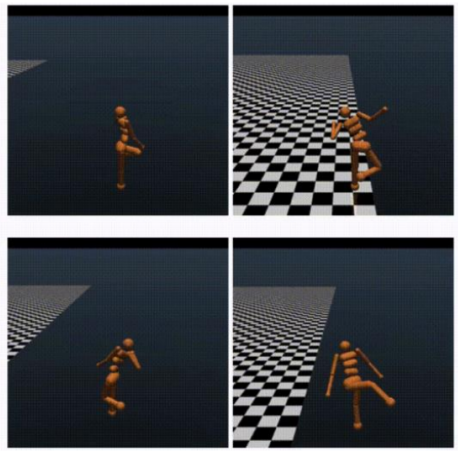


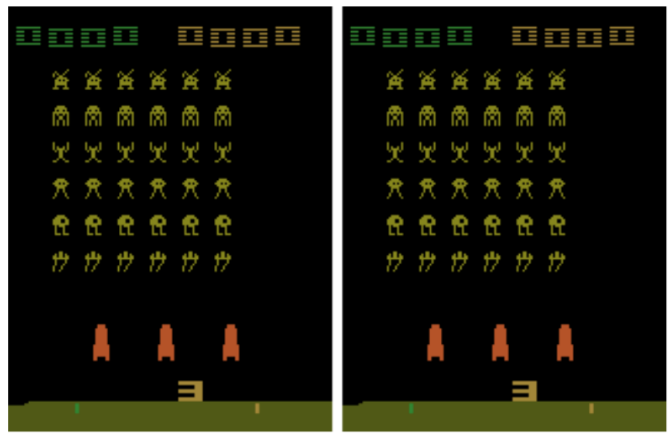
ML branches

- Supervised
  - Classification
  - Regression
- Unsupervised
  - Clustering
  - Anomaly detection
  - Generative models
  - Dimensionality reduction
  - ...
- Reinforcement Learning

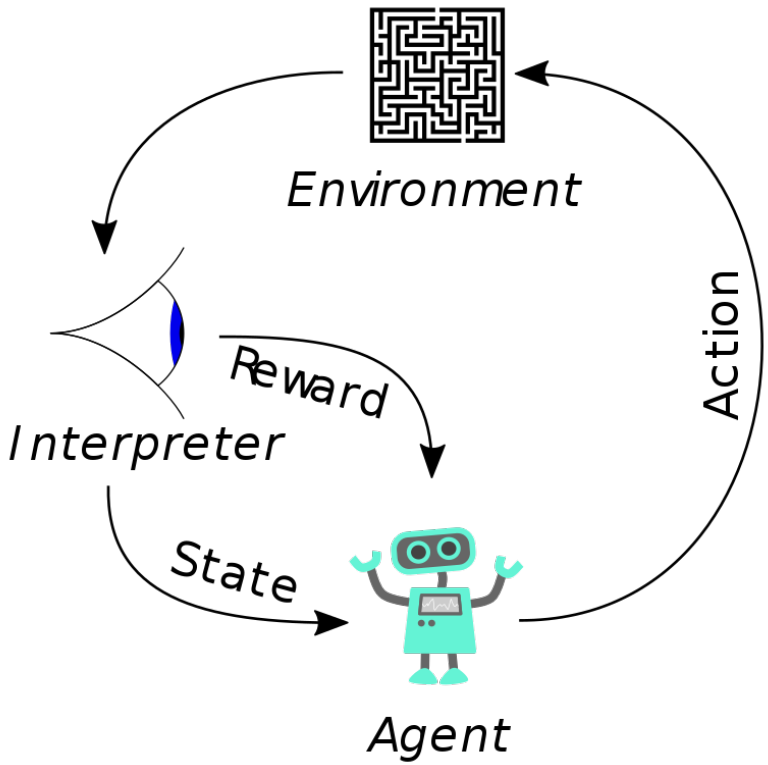
Reinforcement learning



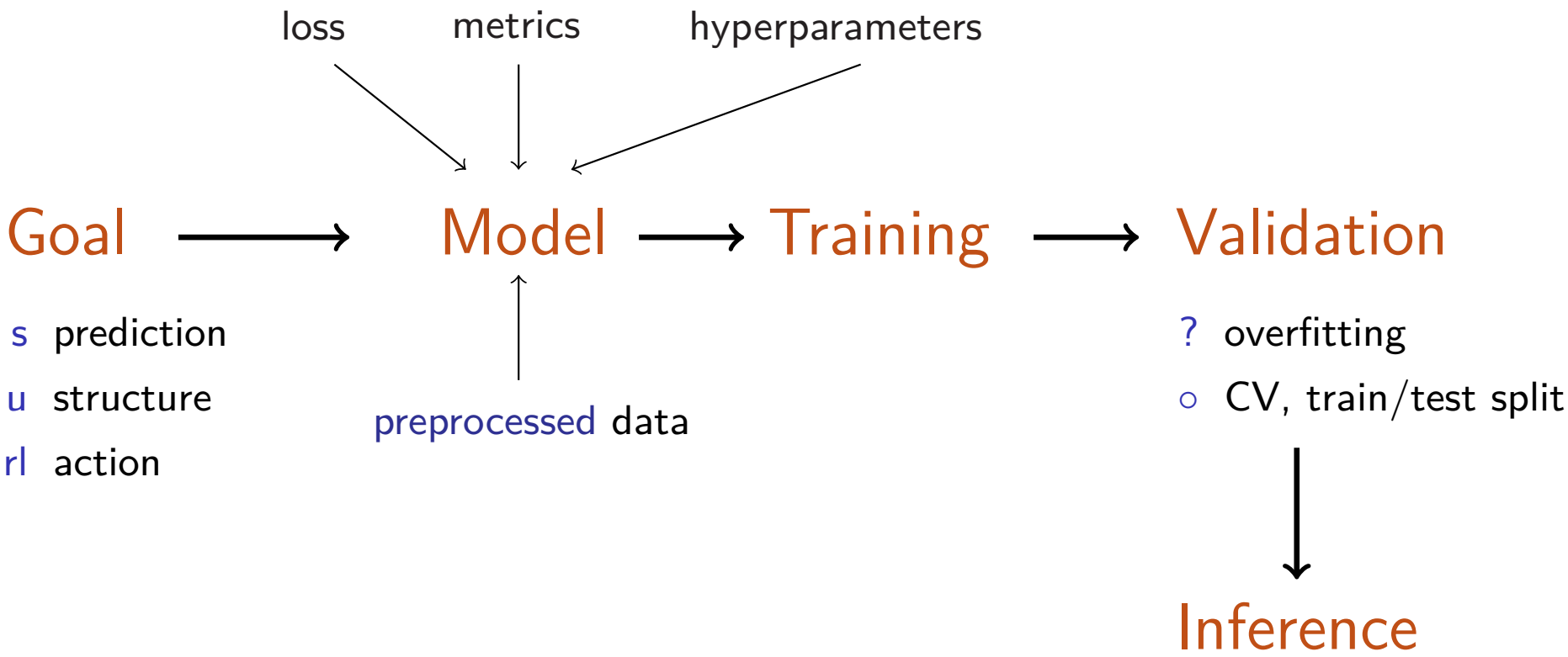




Reinforcement learning



Pipeline

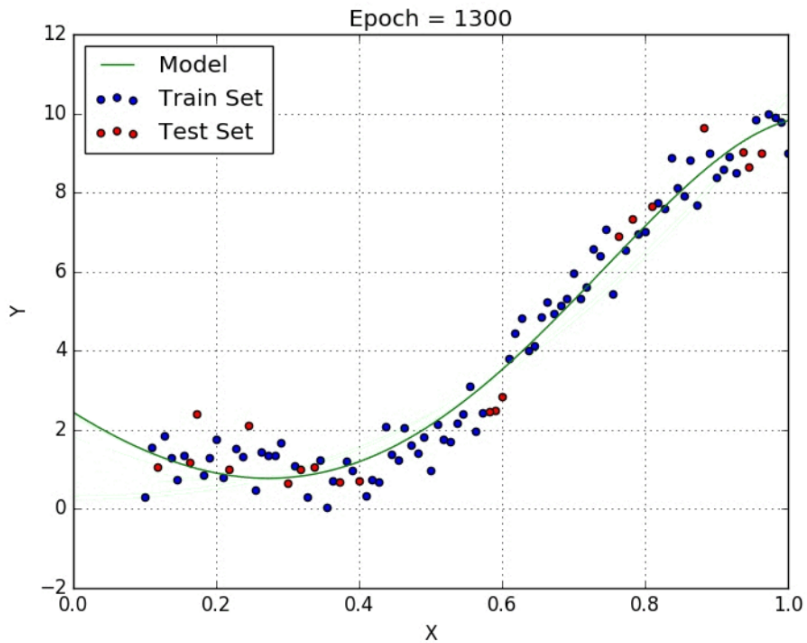

loss
metrics
hyperparameters
Goal
Model
Training
Validation
s prediction
u structure
rl action
preprocessed data
? overfitting
○ CV, train/test split
Inference

Linear models

Linear models

Regression

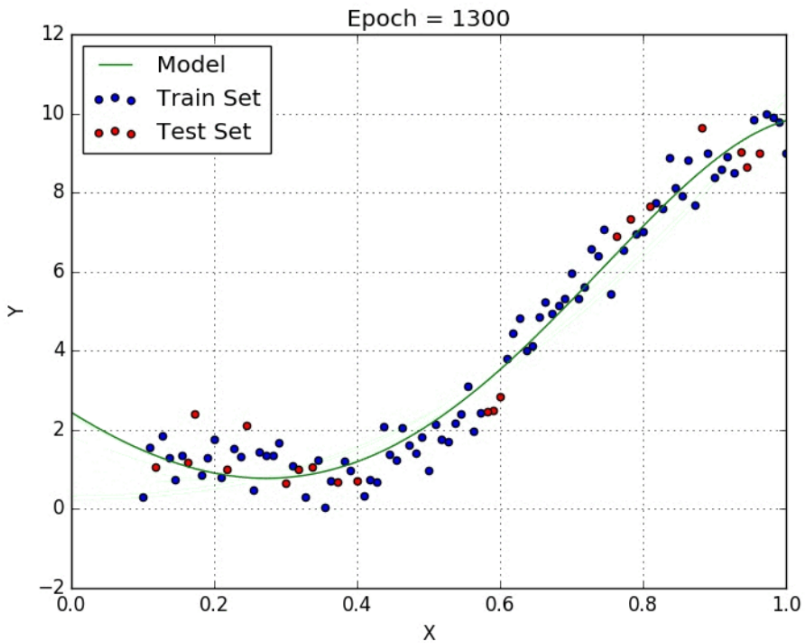
- $\mathbb{Y} = \mathbb{R}^d$
- $a(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$
- L(a,Y):
  - MSE (L2) $= \frac{1}{\ell}\sum_{i=1}^{\ell}(y_i - a(\mathbf{x}_i))^2$
  - RMSE $= \sqrt{\frac{1}{\ell}\sum_{i=1}^{\ell}(y_i - a(\mathbf{x}_i))^2}$
  - MAE (L1) $= \frac{1}{\ell}\sum_{i=1}^{\ell}|y_i - a(\mathbf{x}_i)|$
  - $R^2 = 1 - \frac{\sum_{i=1}^{\ell}(y_i - a(\mathbf{x}_i))^2}{\sum_{i=1}^{\ell}(y_i - \overline{y})^2}$
- $L(a,Y) \to \min_{\mathbf{w}}$



Linear models

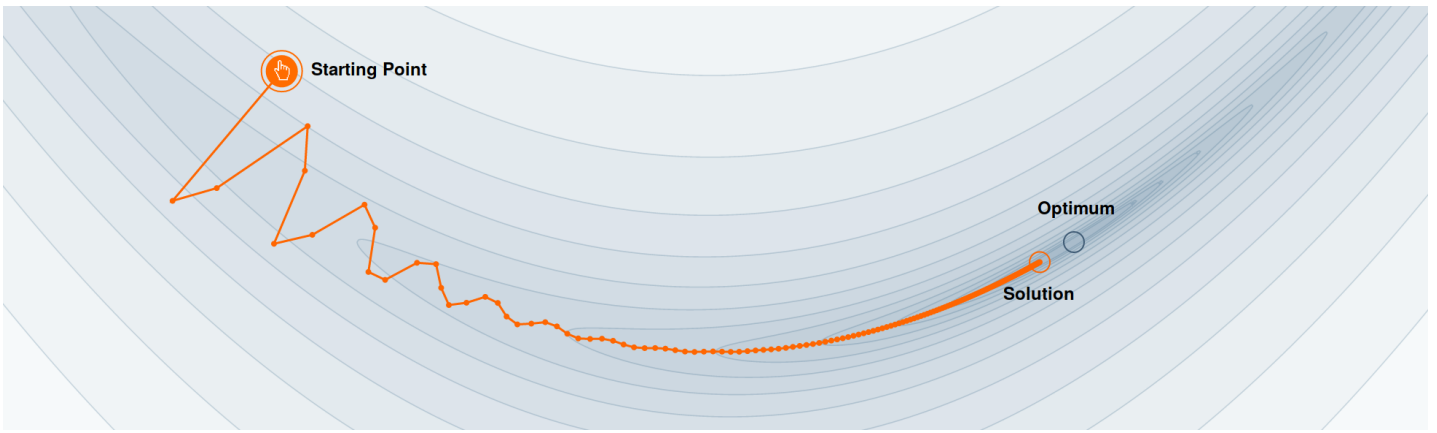
Regression

- $\mathbb{Y} = \mathbb{R}^d$
- $a(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$
- L(a,Y):
  - MSE (L2) $= \frac{1}{\ell}\sum_{i=1}^{\ell}(y_i - a(\mathbf{x}_i))^2$
  - RMSE $= \sqrt{\frac{1}{\ell}\sum_{i=1}^{\ell}(y_i - a(\mathbf{x}_i))^2}$
  - MAE (L1) $= \frac{1}{\ell}\sum_{i=1}^{\ell}|y_i - a(\mathbf{x}_i)|$
  - $R^2 = 1 - \frac{\sum_{i=1}^{\ell}(y_i - a(\mathbf{x}_i))^2}{\sum_{i=1}^{\ell}(y_i - \overline{y})^2}$
- $L(a,Y) \to \min_{\mathbf{w}}$ $\Longleftarrow$ how can we do that?



Linear models

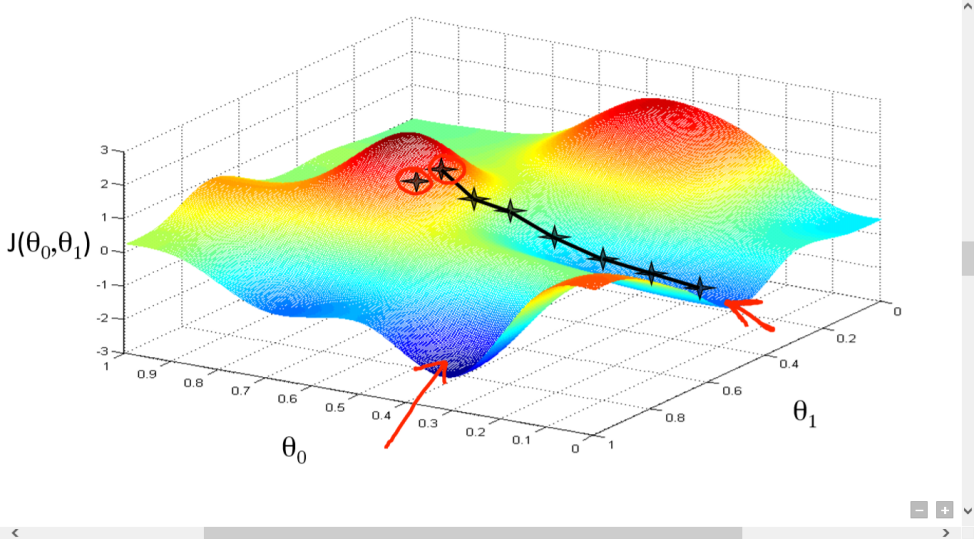
Gradient descent



? $f(\mathbf{x}) \to \min\limits_{\mathbf{x}}$

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_d}\right)$$

! $\mathbf{x}^{(k)} \leftarrow \mathbf{x}^{(k-1)} - \alpha \nabla_{\mathbf{x}} f(\mathbf{x}^{(k-1)})$



more fancy [1], [2]

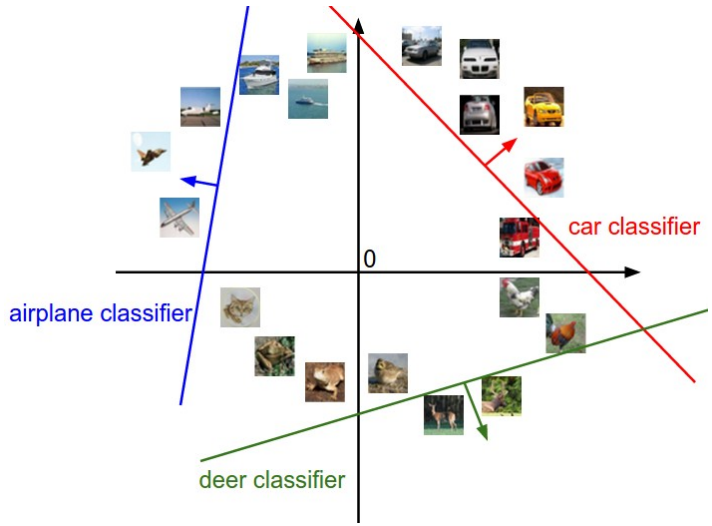
Linear models

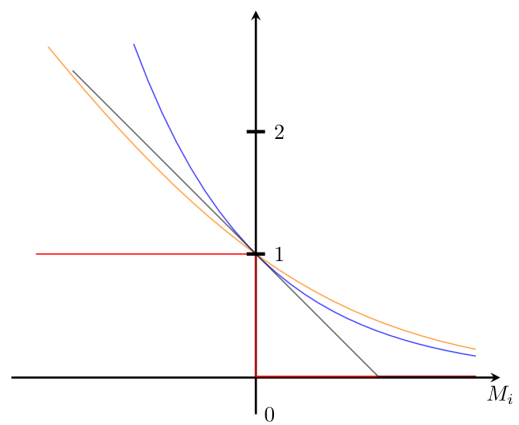
Classification

- $\mathbb{Y} = \{-1, 1\}$
- $a(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x})$
- L(a,Y):
  - Accuracy $= \frac{1}{\ell}\sum_{i=1}^{\ell}[y_i = \text{sign}(\mathbf{w}^T\mathbf{x})]$

? $\frac{1}{\ell}\sum_{i=1}^{\ell}[y_i = a(\mathbf{x}_i)] \to \min_{\mathbf{w}}$

! Optimize upper bound $\tilde{L}(a,Y)$
  - Log loss
  - Hinge loss
  - ...



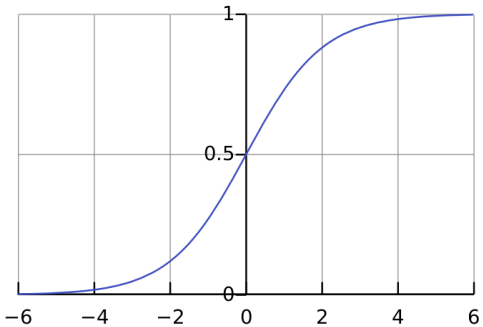


and more fancy [1], [2]

Linear models

Logistic regression

$\mathbb{Y} = \{-1, +1\}$



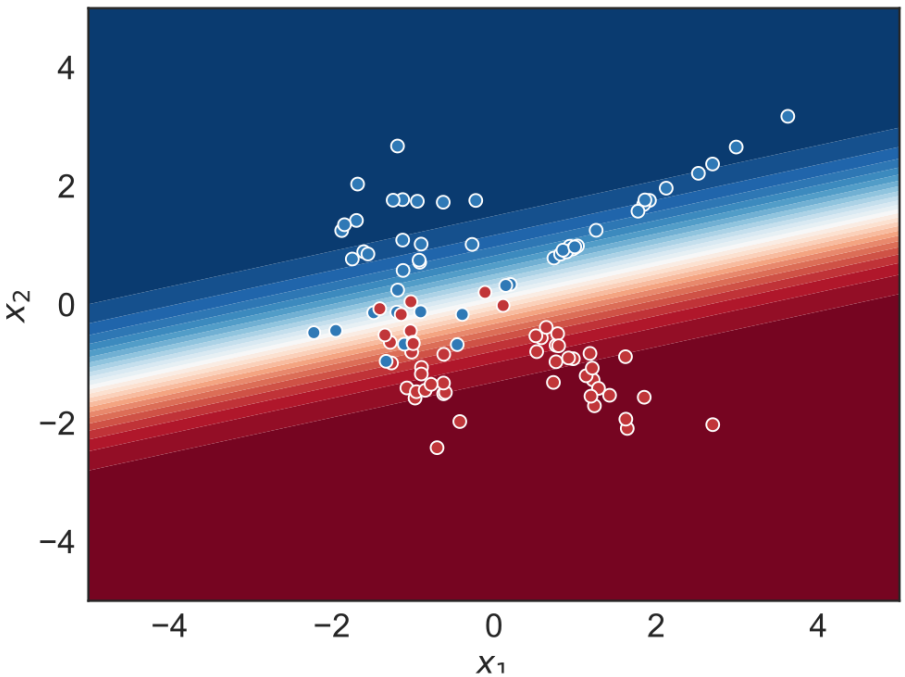
- $a(\mathbf{x}) = \dfrac{1}{1 + \exp(-\mathbf{w}^T\mathbf{x})} \equiv \sigma(\mathbf{w}^T\mathbf{x})$
- L(a,Y):
  - Probability: $a(\mathbf{x}_i)^{[y_i=+1]} \cdot (1 - a(\mathbf{x}_i))^{[y_i=-1]}$
  - Likelihood: $\mathcal{L} = \prod_{i=1}^{\ell} a(\mathbf{x}_i)^{[y_i=+1]} \cdot (1 - a(\mathbf{x}_i)^{[y_i=-1]}$
  - Log likelihood: $\log \mathcal{L} = \sum_{i=1}^{\ell} [y_i = +1] \log a(\mathbf{x}_i) + [y_i = -1] \log(1 - a(\mathbf{x}_i))$

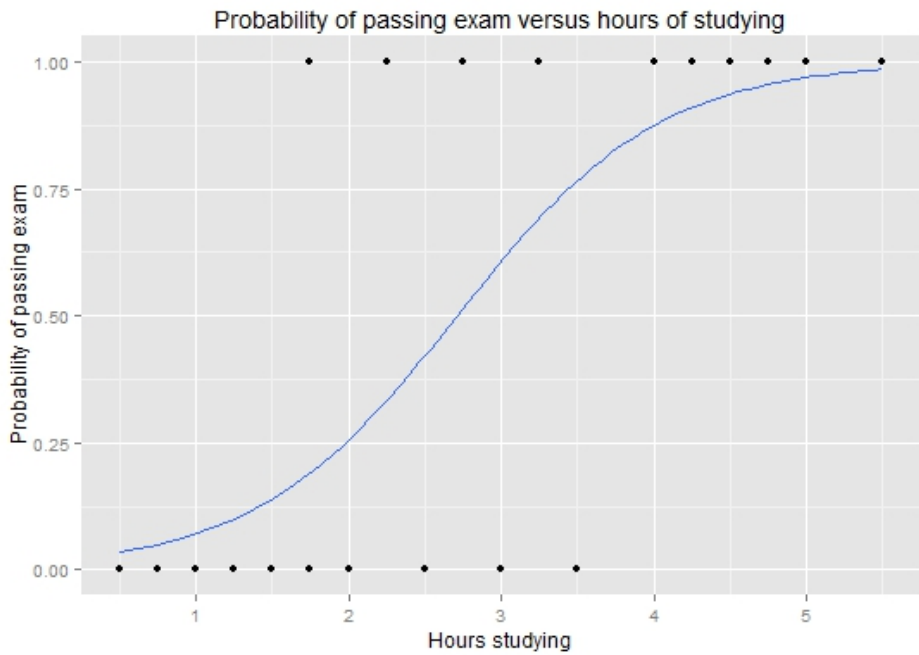
! predicts $P(y = +1|\mathbf{x})$

! interpretable $\mathbf{w}$ (in general for linear models)

Linear models

Logistic regression





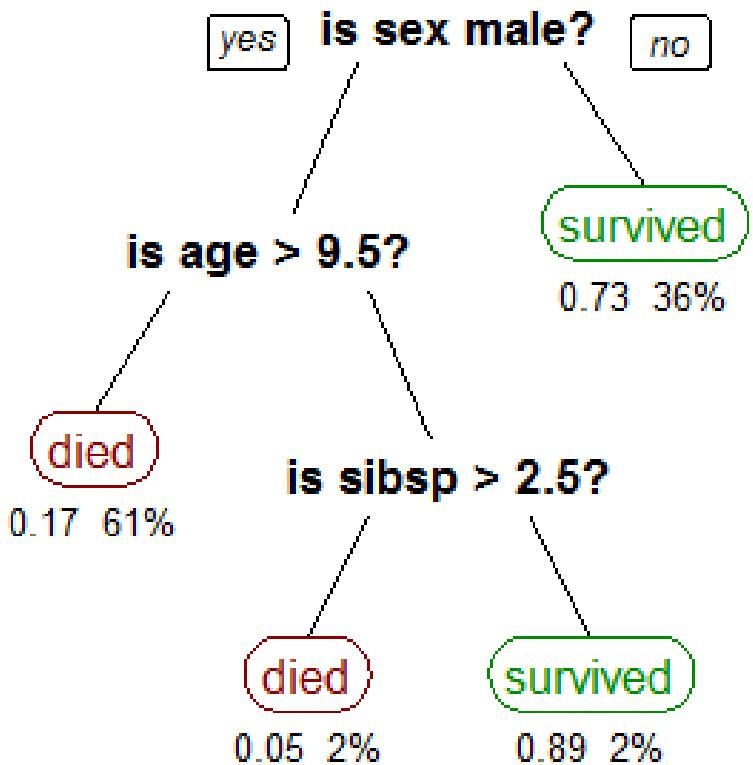
Linear models

pros & cons

+ Interpretable
+ Unique solution exists
+ Nice statistical properties
+ Fast
+ Perform well with lots of features
+ OK with little data
- Don't capture non-linear correlations
- Require feature linear independence

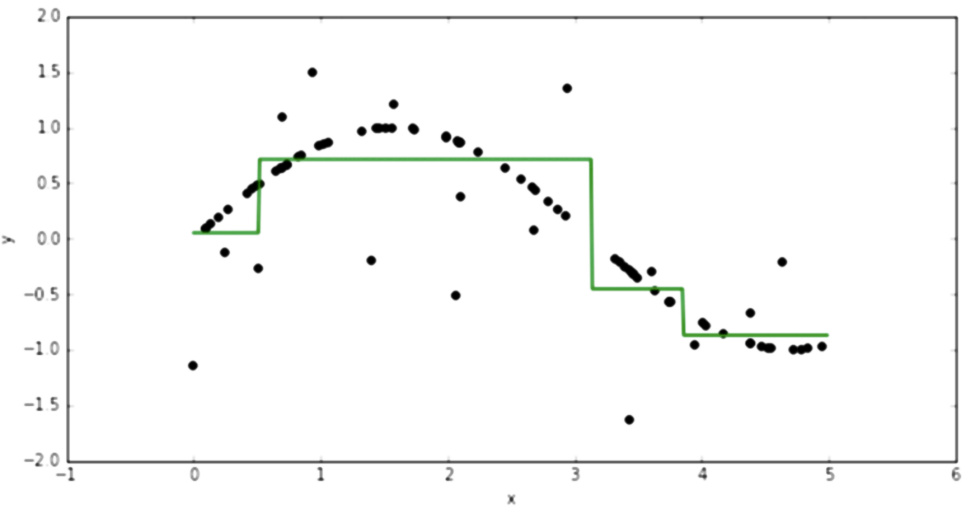
Trees

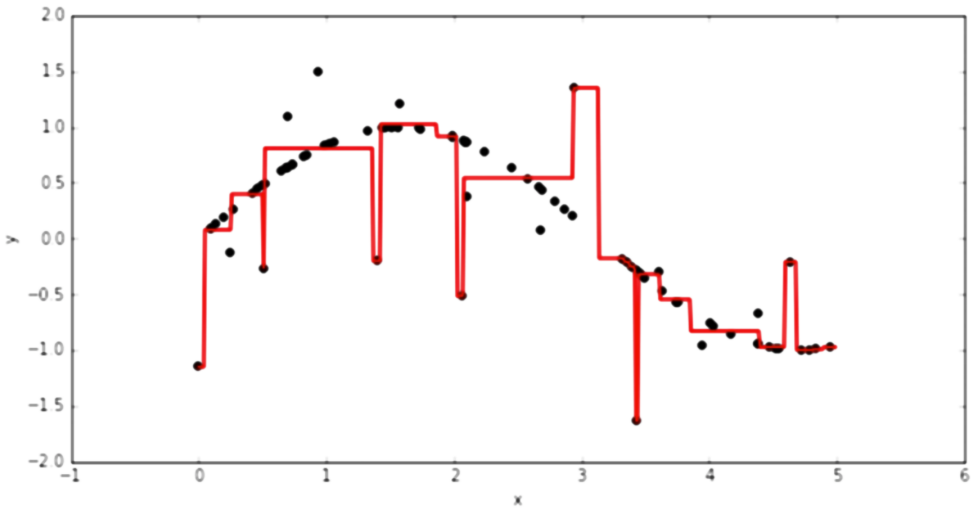
Trees


yes
is sex male?
no
is age > 9.5?
survived
0.73 36%
died
0.17 61%
is sibsp > 2.5?
died
0.05 2%
survived
0.89 2%

Trees

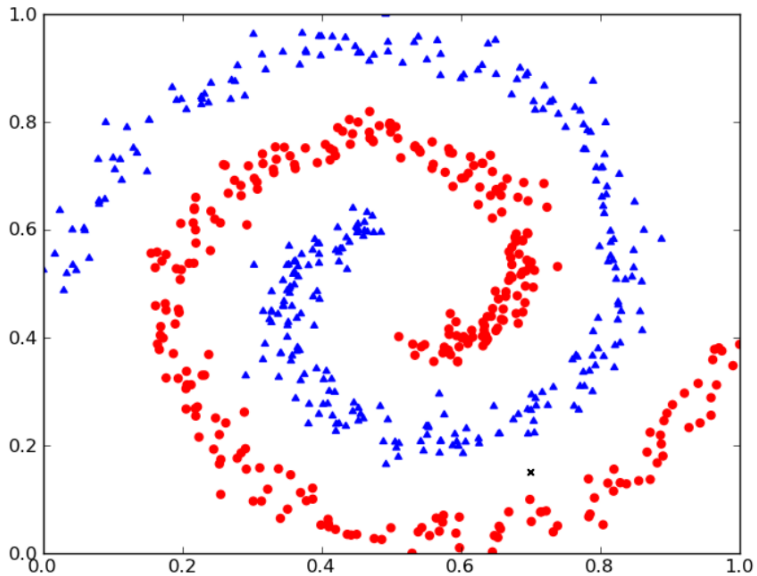
Regression

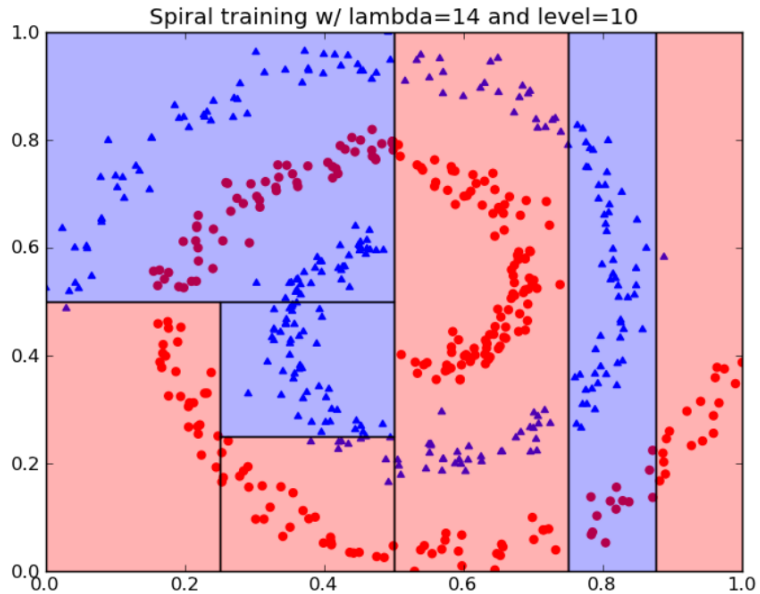




Trees

Classification



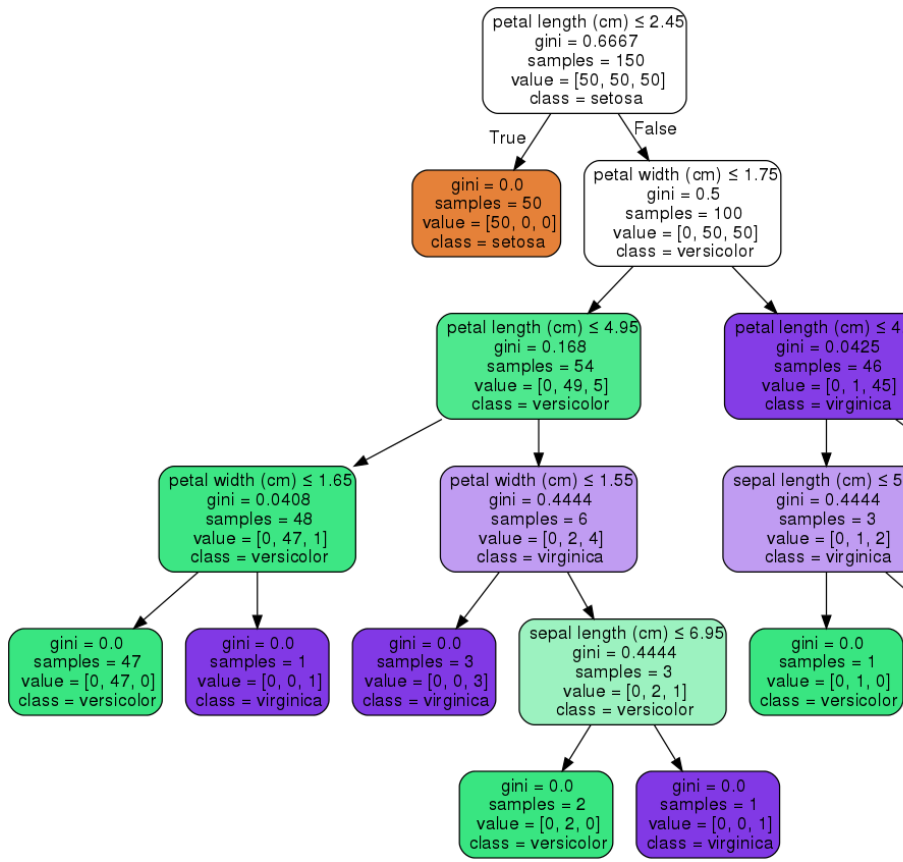


awesome visualisation [link]

Trees

Training

- Greedily split into Q_ℓ and Q_r:
  - $Q_\ell(\theta) = (x,y)|\, x_j \leq t_m;\ \theta = (j, t_m)$
  - $Q_r = Q \backslash Q_\ell(\theta)$
- $G(Q,\theta) = \frac{n_\ell}{N} H(Q_\ell(\theta)) + \frac{n_r}{N} H(Q_r(\theta))$
- Impurity $H(Q)$:
  - c Gini
  - c Entropy
  - r MSE (target variance), MAE
- minimize $\theta^* = \text{argmin}_\theta G(Q,\theta)$
- stopping criteria
- leaf prediction - voting/mean value

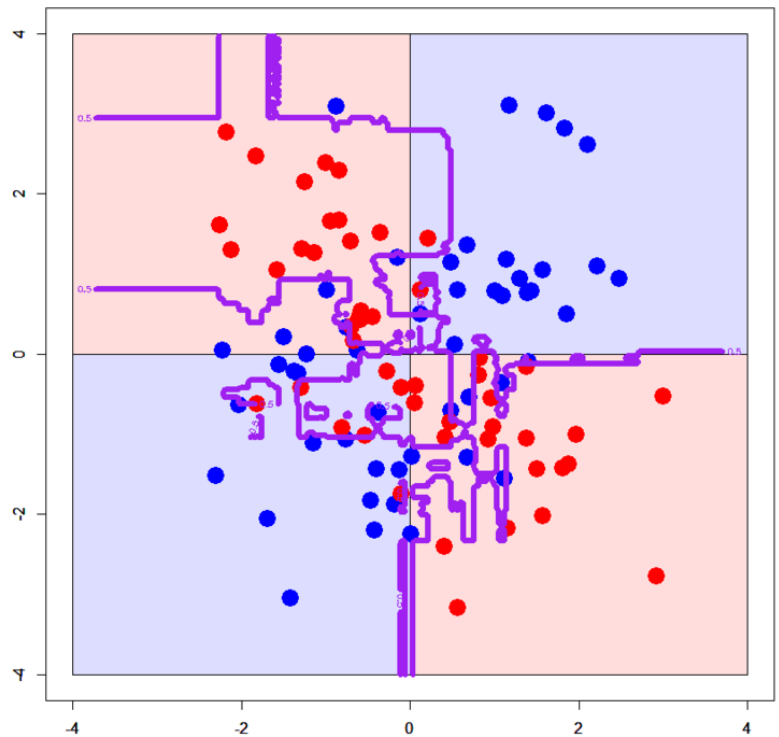


further reading [1], [2]

Trees

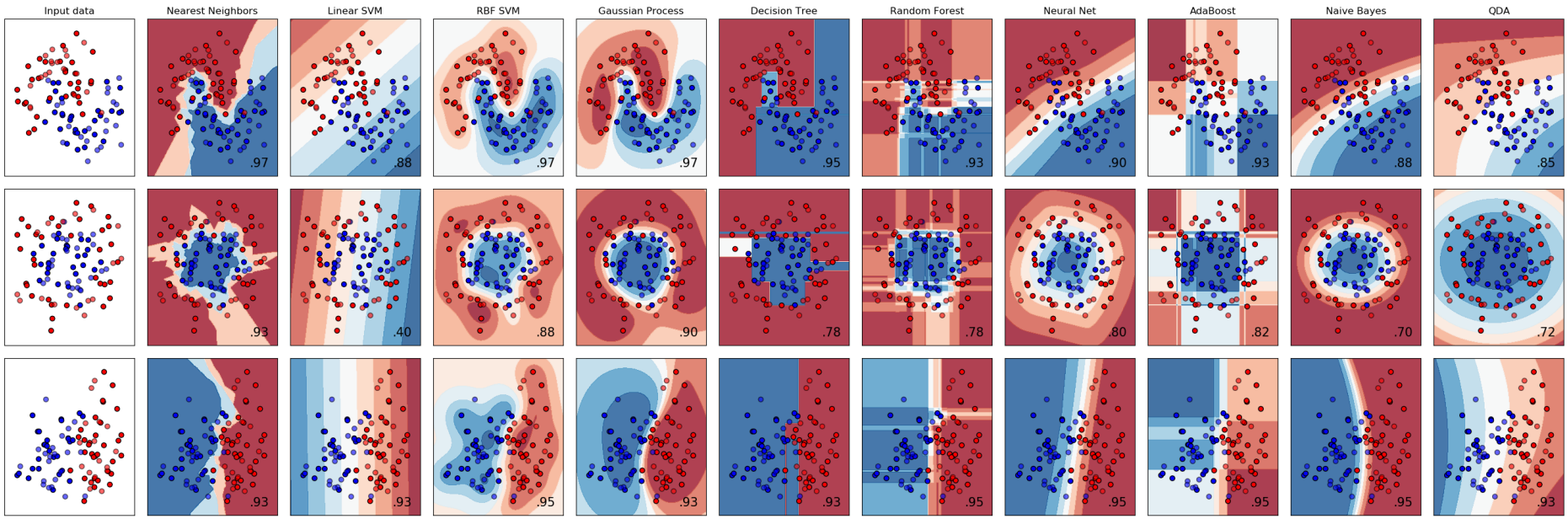
pros & cons

+ Interpretable
+ Categorical features
+ Feature linear dependence OK
- Overfits easily
- Not that powerful



comparison here [link]

there's more!



Out of the scope: Naive Bayes, kNN, SVM

ML specifics

Metrics

They are really important

loss function

model validation

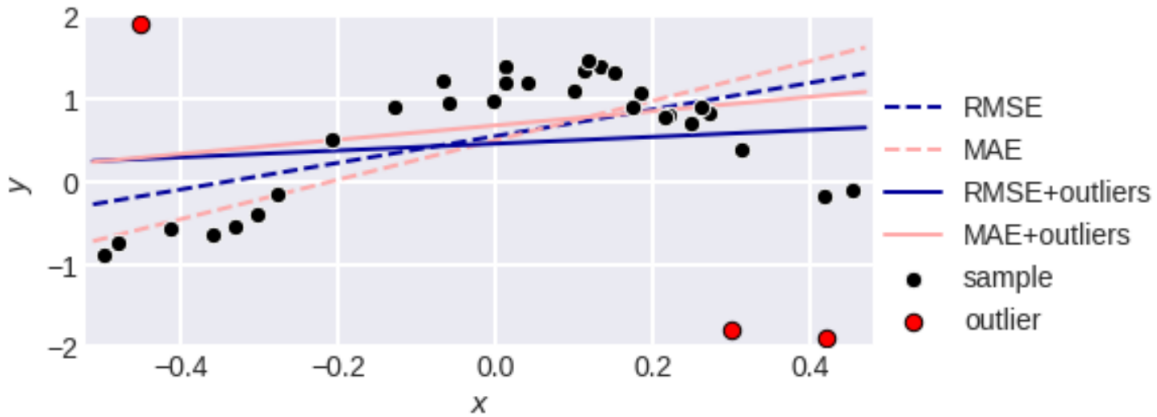
hyperparameter optimization

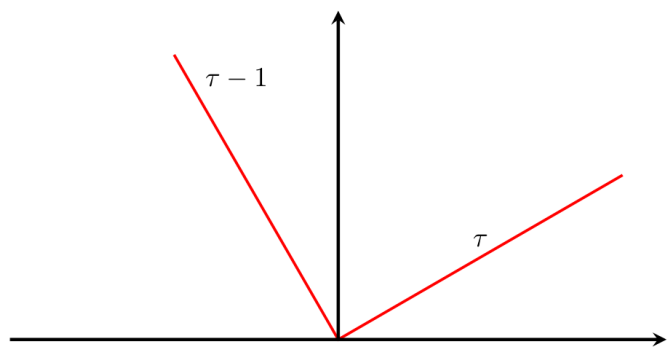
Metrics

Regression

nice one on metrics [1], [2]

- MSE $= \frac{1}{\ell}\sum_{i=1}^{\ell}(y_i - a(\mathbf{x}_i))^2$
  - sensitive to outliers
  - easy to optimize
- MAE $= \frac{1}{\ell}\sum_{i=1}^{\ell}|y_i - a(\mathbf{x}_i)|$
  - robust to outliers
  - gradient problems (no)
- $R^2 = 1 - \dfrac{\sum_{i=1}^{\ell}(y_i - a(\mathbf{x}_i))^2}{\sum_{i=1}^{\ell}(y_i - \overline{y})^2}$
  - explained variance
  - comparison with naive mean model
- Logloss





Metrics

Classification

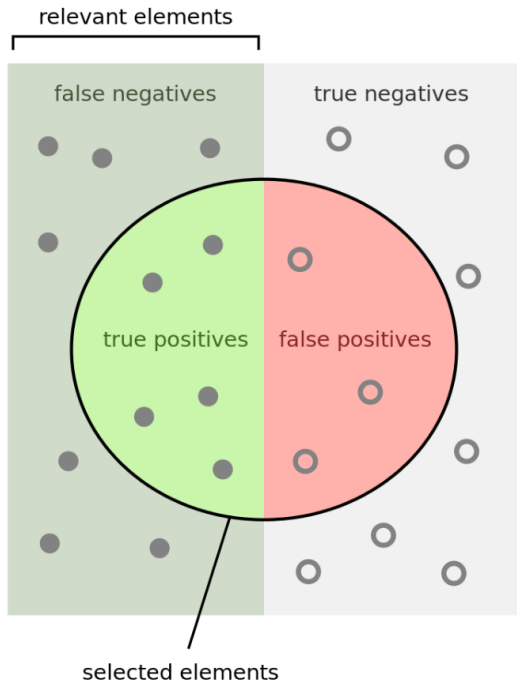
further reading [link]

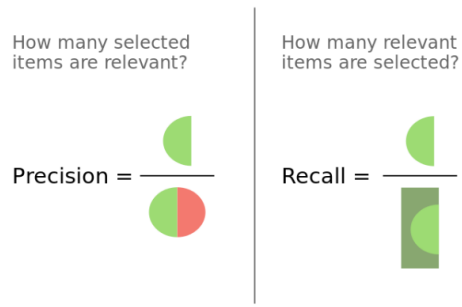
- Accuracy $= \frac{1}{\ell}\sum_{i=1}^{\ell}[y_i = a(\mathbf{x}_i)]$
  - imbalanced data: 990b vs 10s
  - equal error weight
- Confusion matrix
  + more info
- Precision, Recall
- F-score $= \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$

! they are all threshold dependent

- Logloss

	$y = 1$	$y = -1$
$a(x) = 1$	True Positive (TP)	False Positive (FP)
$a(x) = -1$	False Negative (FN)	True Negative (TN)

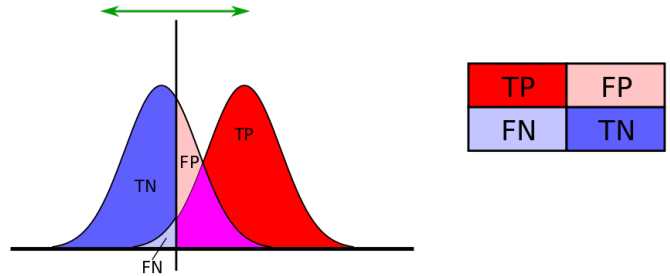


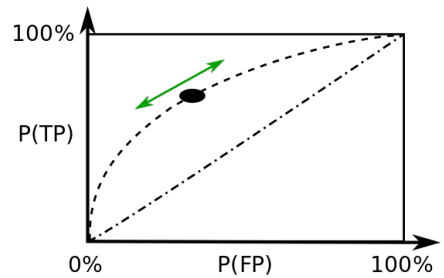


cool interactive demo [1]
more facts [2]

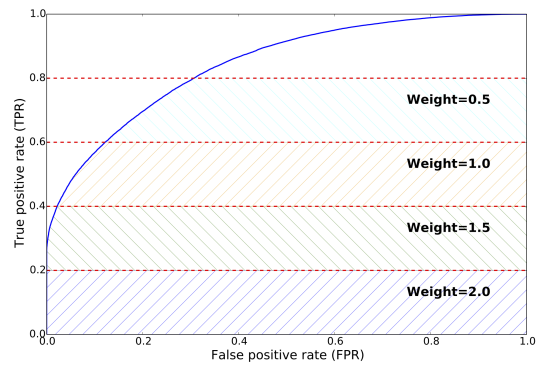
Metrics

ROC curve

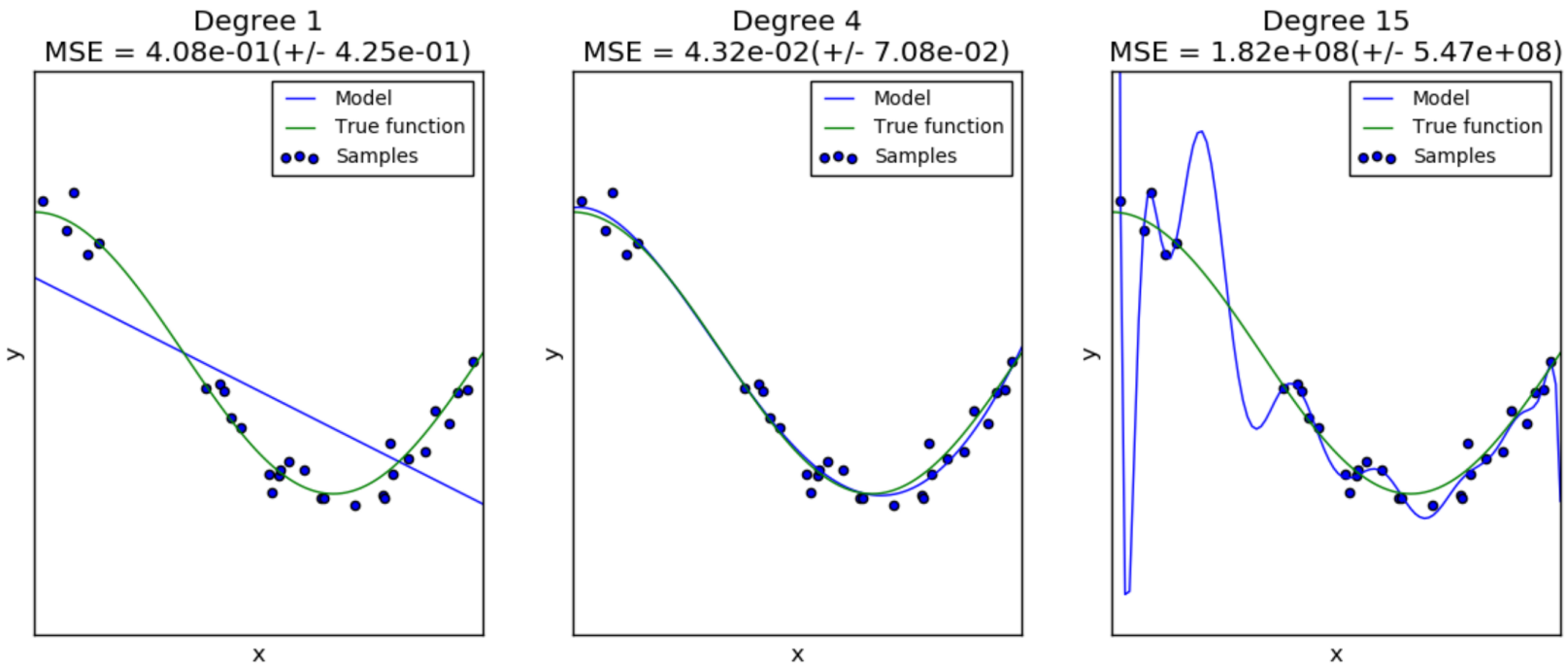




- curve in (TPR(t), FPR(t)) space
  corr. to different thresholds $a(\mathbf{x}) > t$
  - $\mathrm{TPR} = \dfrac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$
  - $\mathrm{FPR} = \dfrac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$
- ROC AUC = Area Under Curve
  - describe classifier as a whole
  - class separation (but not proba precision)
  - probability of correct pairs ordering
- weighted ROC AUC [3]



Overfitting

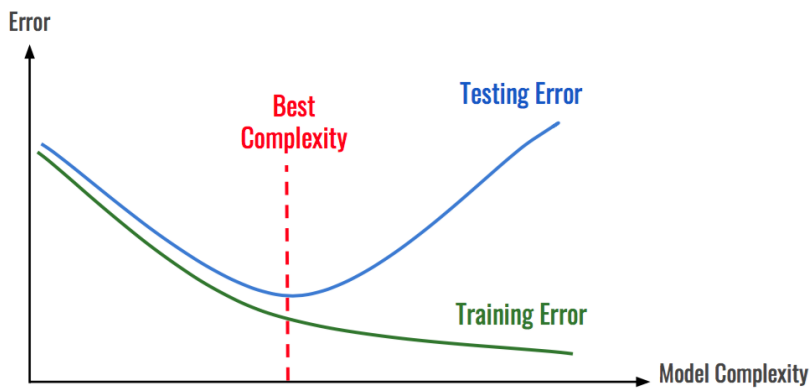


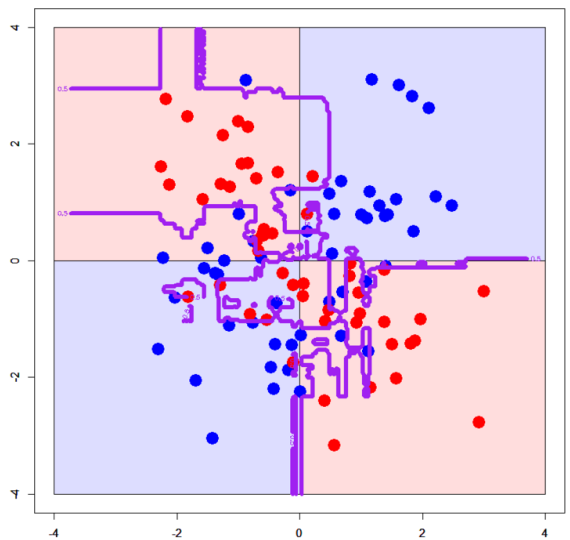
Overfitting

- Often looks like
  - $a(\mathbf{x}) = 0.5 + 12458922x + 43983740x^2 + ... + 2740x^9$

  $\rightarrow$ regularization
- Generalization:
  - underfitting - bad on train, bad on unseen
  - overfitting - excellent on train, bad on unseen

  $\rightarrow$ validate on hold-out data





Regularization

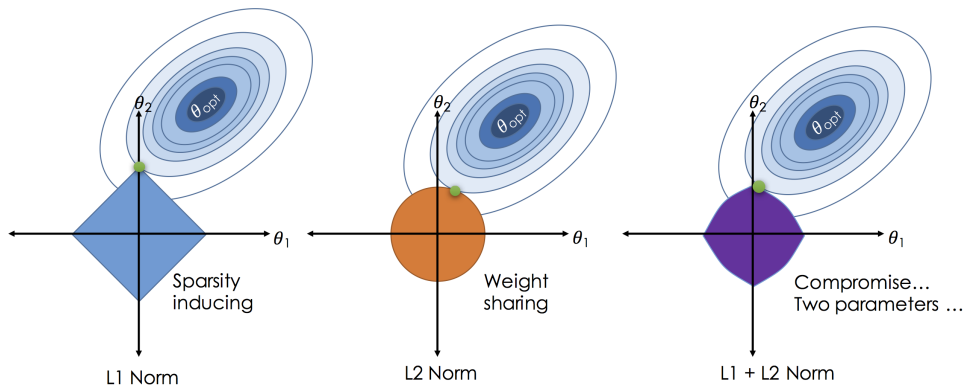
- Add penalty term to loss function:

  L1: $||w||_1 = \sum_i |w_i|$

  L2: $||w||_2 = \sum_i w_i^2$

- $L(w, X) + \lambda||w||_{1,2} \to \min_w$
- λ is a hyperparameter of the model

! L1 zeros redundant w



Model validation

? check overfitting

? estimate model performance

? optimize hyperparameters

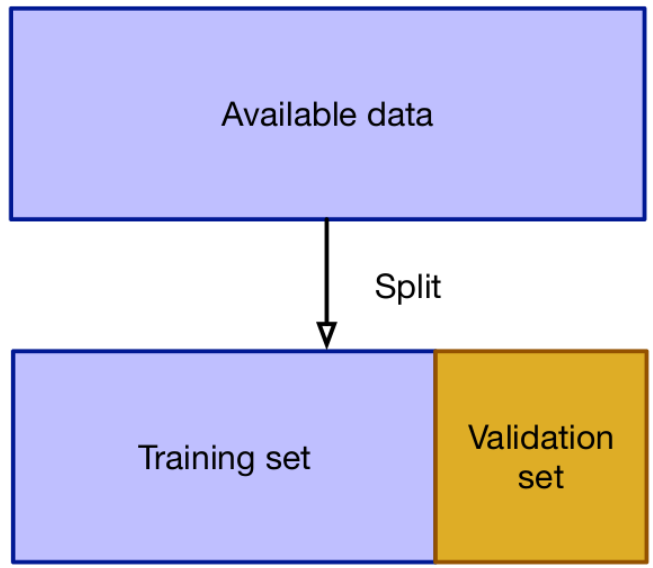
Solution

- train/test split
- cross validation (CV)

Model validation

train/test

- might be luxury
- suffer from bad splitting
- you may overfit



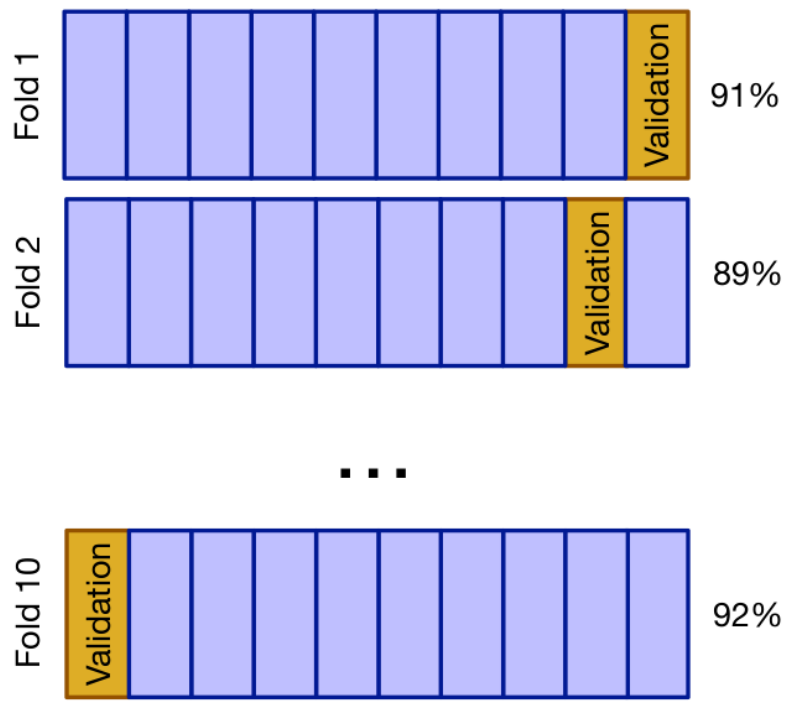
Model validation

CV

- $\text{CV} = \frac{1}{K}\sum_{i=1}^{K} Q(a, X_k)$

+ sensitive to overfitting
+ can use all the data
- still may suffer from bad splitting
- many folds
  + small bias
  - large variance
  - time-consuming
- few folds
  + small variance
  + computationally effective
  - large bias
- often used for hyperparameter optim.
  - e.g., polynomial degree d

→ train/test(CV)/val split




#лидируйвалидируй

Summary

- What is ML
  - supervised [s]
  - unsupervised [u]
  - reinforcement learning [rl]
- Gently into [s]
  - linear models
  - trees
- ML specifics
  - metrics
  - overfitting
  - regularization
  - model validation

next lecture

- Tree ensembles
- Feedforward NN
  - CNN & RNN