pg 41 03/03/21

Ch. 2 Notes

(1) Big Picture
(2) Frame the problem → Pipelines
(3) ~~Reference &~~ Notation & Performance Metrics

$x^{(i)}$ is the ith feature vector
eg. If 4 ~~predictors~~ attributes in a given feature labeled x_n

Then $(x^{(1)})^T = (x_{11}, x_{21}, x_{31}, x_{41}) = (1000, 423, 322{,}000, 4)$
random #s

$X =$ matrix of features

$$\begin{pmatrix} (x^{(1)})^T = (x_{11}, x_{12} \dots) \\ \vdots \\ (x^{(i)})^T = (x_{i1}, x_{i2} \dots) \end{pmatrix}$$

h ← hypothesis ~~fxn~~ (prediction fxn)
$\hat{y}^{(i)}$ is a predicted value

- RMSE is sensitive to outliers (Euclidean ~~norm~~ norm, or ℓ_2 or $\|\cdot\|_2$ or $\|\cdot\|$)
- MAE uses Manhattan norm

$$MAE(X, h) = \frac{1}{m} \sum_{i=1}^{m} |h(x^{(i)}) - y^{(i)}|$$

(4) Check Assumptions
— what will data be used for what, are assumptions of models, etc.

pg 39 **Note** Univariate regression means predict ONE $y^{(i)}$
Multivariate regression means predict a set(?) of $y^{(i)}$

pg 2

~~(5) Test set~~

(5) look @ data structure
-- info( ) # Num nonnull & dtypes
-- describe( ) # estimators
-- value_counts( ) # Individual instances
-- hist(bins=..., figsize=(,))

(6) Create a test set
-- don't look to avoid DATA SNOOPING
-- avoid sampling bias by stratified samples

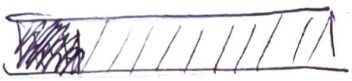
youtube

Types of CV

@ Random states

① ~~Leave one out~~ Leave P out CV

Obsolete

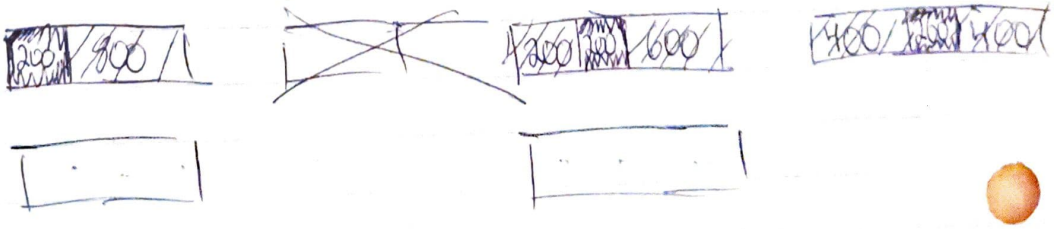
exp1  Test Train

exp1 

notes: • Requires many iterations to be representative
• how bias: you get better @ only 1 model

② k-fold CV

k=5 (5 experiments)
+ folds are m ∴ with 1000=m + k=5, each fold consists of k of 200 ~~data~~ units (?)



Final mean of ~~each~~ accuracy across k-folds or tell other estimators (min, max, mean)

Pg 3

(2) K-Fold - CV

Notes:

- If clustering of data (eg. Think classification) Then you get unrepresentative results

(3) Stratified CV

- At least some # of each class is contained w/i training & testing
- Dependent on proportions in data set

☆ (4) Time-series cross validation

T1
T2
T3
T4
T5
T6 ← Targets
T7 ←

Predict stock eg:
~~D1 D2 D3 D4 D5~~

Stock					Output
D1	D2	D3	D4	D5	D6
D2	D3	D4	D5	D6	D7

Pg 4

- Encoding:
  -- one hot is not ideal for categories w/ large #'s of attr

  e.g. professions ⇒ Cook, Police man, etc.
  Country code ⇒ ... ...
  species ⇒ ~~1~~ animal 1, two, 3, ...

- Feature Scaling
  (1) Min-max scaling ← Sklearn MinMaxScaler
  (2) Standardization
      ↳ Typical statistical procedure that accounts for $\sqrt{\sigma} = s_x$?
  -- does not bound to specific range
  -- less affected by outliers

★ ~~WHERE TO APPLY TRANSFORMATIONS~~
Apply only on ~~test~~ ~~valid~~ holdout validation set @ very end. Otherwise it's fine to reuse this.