HW 2 notes

Data splitting, RoC

Loose ends from HW2

- A majority class baseline
  - Powerful if one class dominates. Often happens in real life
  - Recognizer becomes biased towards the majority class (the prior term)
  - How to deal with this?
- Zero probability in the estimation
- Hyperparameter
- RoC

Splitting the data

- We want to estimate the performance on the model
  - Need a test set

- Train-test split
- Leave-one-out splitting
- Bootstraping
- Cross-validation splitting

Splitting data

Simple train-test split

Ratio can be varied as appropriate

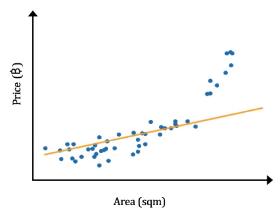
Stratified splitting – tries to keep the distribution in the training and test the same

sklearn.model_selection.StratifiedShuffleSplit
sklearn.model_selection.StratifiedKFold

Leave-one-out

Multiple train-test split

Estimate the true performance (expected performance)



Random

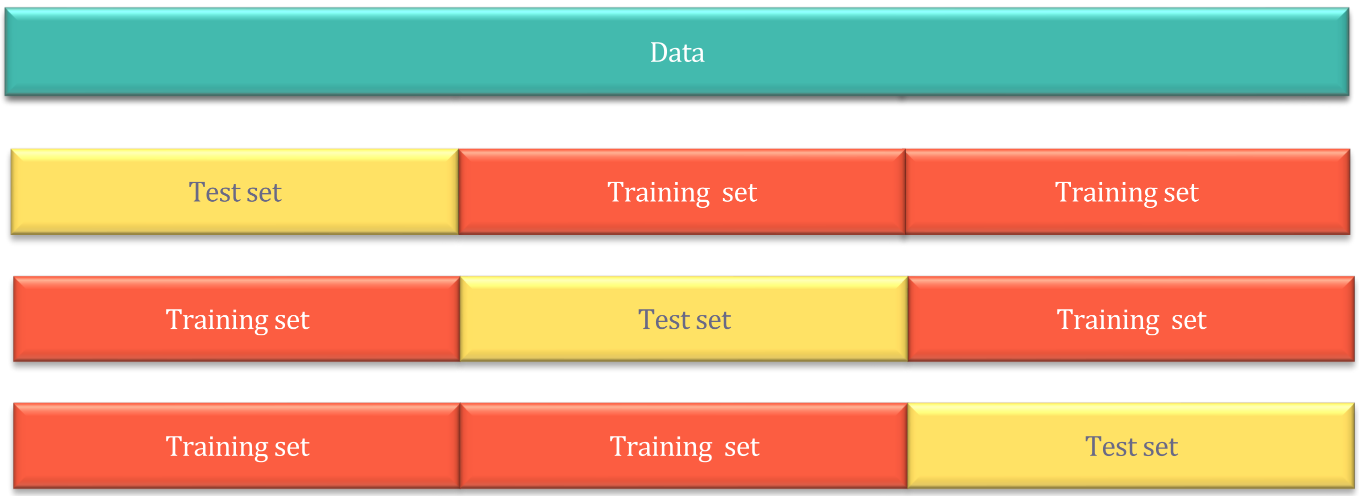
$\hat{y} = 2.3x1 - 3.4x2 + 4.2$

Train-test split → Train model → Predict

x 10

This is sometimes called bootstrapping (in statistics).
This can also be used to calculate the variance of your method performance

Cross-validation (CV)



3-fold cross-validation

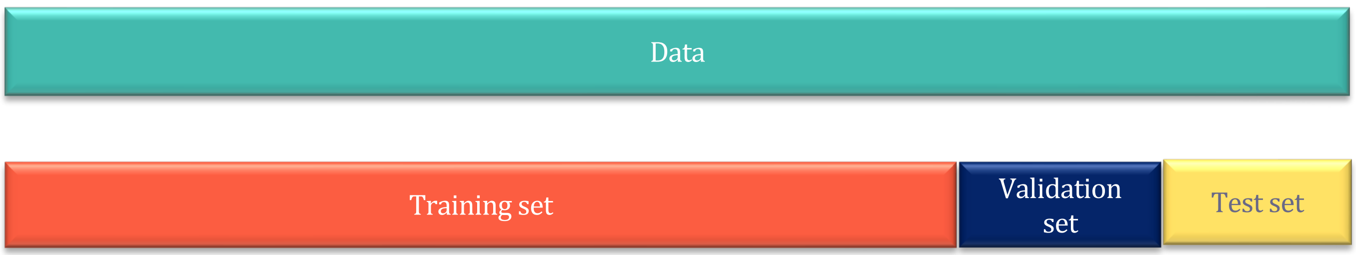
Similar idea to bootstrapping but there's no overlapped in the splits.

Hyperparameter vs Parameters

- Parameters - something the models learn from data
- Hyperparameter – something we pick for the model via trial and error

Model	Parameter	Hyperparameter
Linear regression	weights	Loss (L1/L2), polynomial degree, features used, …
Naïve Bayes	Distribution parameter	Type of distribution used
GMM distribution	Means, covariance, mixture weight	Number of mixture, type of covariance matrix
Histrogram distribution	Histrogram height	Number of bins or size of bins
K-means	Centroids	K
ML model	Model weights	Type of ML model

Picking hyper-parameters

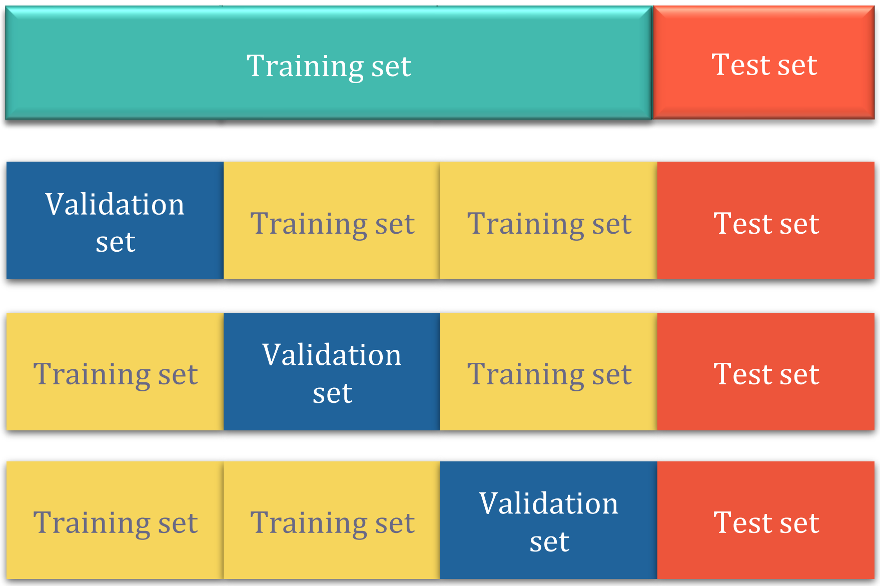


You decide on the models and hyperparameter on the validation performance.

Make sure that you are not optimizing on the test set.
Make the test set a good proxy for estimating real world performance

Don’t cheat!

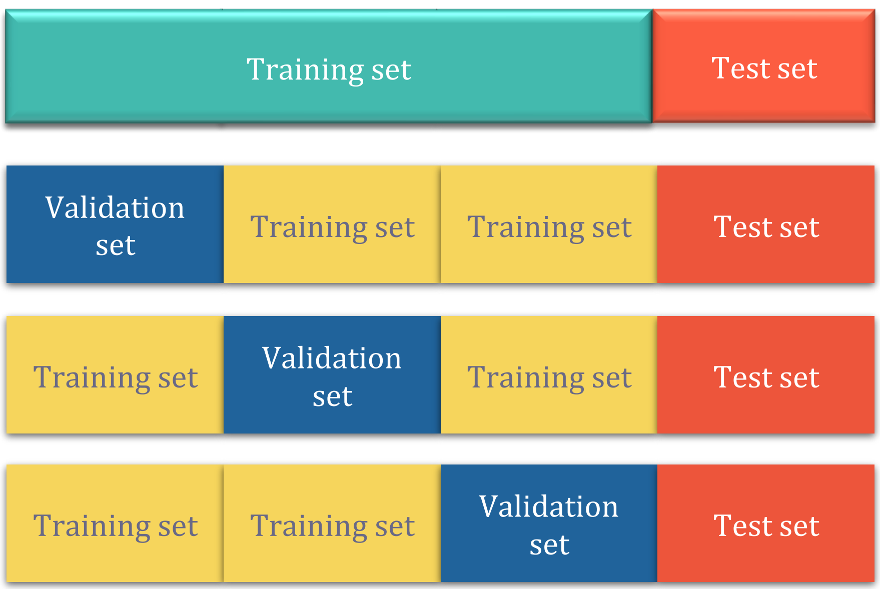
Splitting a validation set



If the test set is fixed, we can do CV on the training set to get the valitation set

Estimating true performance of our pipeline

Freeze the test set
Touch the test set as less as possible
The more you often see the test set the more you cheat



Example

Best degree = 3, CV accuracy = 80, Test accuracy = 75

Best degree = 2, CV accuracy = 78, Test accuracy = 80

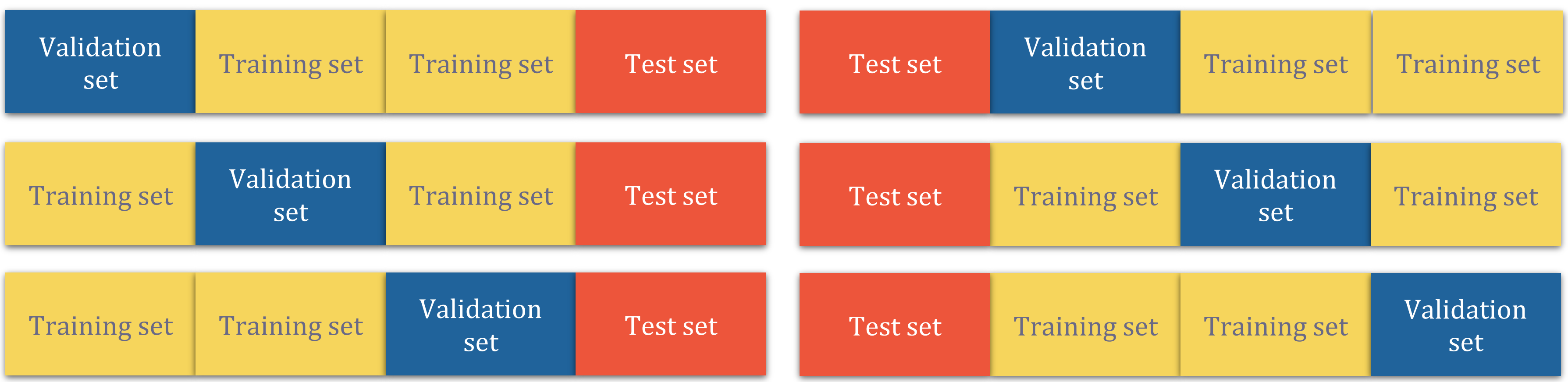
Best degree = 2, CV accuracy = 78, Test accuracy = 85

Estimates of the accuracy using our precedure = 80%

Question: which model do we deploy?

Nested CV

It the fixed test set is too small to give a reliable estimate, use nested CV.
Or a mixture of tecchniques, EX leave one out CV with a validation set.



Test split 1

Test split 2

Size of split

- Train/validation/test split
  - 80/10/10, 90/5/5, 5-fold CV, leave one out CV, etc. for academia
- For real applications, get dev and test sets that represent your users.
  - Reflects the data you want to do well on.
  - There can be a mis-match between train and dev data. But avoid mis-match between dev and test data.
  - If no users, recruit friends to pretend to be the users.

- Example: Cat classifier.
  - Should you use ImageNet cat pictures as train/dev/test?
  - Go pretend you’re a user and take cat pictures for the dev/test set.

Val and test set and size

- Val - tune hyperparameters, select features, and make other decisions regarding the learning algorithm.
- Test - evaluate the performance of the algorithm, but not to make any decisions about regarding what learning algorithm or parameters to use.

- Val – big enough to notice difference between algorithms (if you care about 0.1% difference, make sure you have enough dev set to spot it).
- Test – large enough to give confidence that your model will do well in real task

Congratulations on your first attempt on (almost) re-implementing a research paper!

RESEARCH ARTICLE | OPEN ACCESS

Mining housekeeping genes with a Naive Bayes classifier

Luna De Ferrari and Stuart Aitken

BMC Genomics 2006 7:277 | https://doi.org/10.1186/1471-2164-7-277 |

Received: 28 February 2006 | Accepted: 30 October 2006 | Published: 30 October 2006

The main advantage of Bayesian classifiers is that they are probabilistic models, robust to real data noise and missing values. The Naive Bayes classifier assumes independence of the attributes used in classification but it has been tested on several artificial and real data sets, showing good performances even when strong attribute dependences are present. In addition, the Naive Bayes classifier can outperform other powerful classifiers when the sample size is small. Using the words of Domingos and Pazzani: "In summary, [...] the Bayesian classifier has much broader applicability than previously thought. Since it also has advantages in terms of simplicity, learning speed, classification speed, storage space and incrementality its use should perhaps be considered more often." [18].

Another trick to reduce 0 bins in histograms

Algorithms

Discretisation

The Weka algorithm used for filtering with Unsupervised discretisation involves separating the data in ranges using equal-frequency binning (histogram equalization) so that the same number of training example fall into each bin. No class information is taken into consideration [29]. For

Prediction and thresholds

Beer	Grass	Rice	Flood	Prediction
100	3	3	**Yes**	**0.8**
20	1	1	**Yes**	**0.3**
80	3	2	**No**	**0.6**
40	1	1	**No**	**0.2**
40	1	1	**No**	**0.1**

What happens if I set my threshold at 0.5?

Prediction and thresholds

Beer	Grass	Rice	Flood	Prediction	Metric
100	3	3	**Yes**	**0.8**	**TP**
20	1	1	**Yes**	**0.3**	**FN**
80	3	2	**No**	**0.6**	**FA**
40	1	1	**No**	**0.2**	**TN**
40	1	1	**No**	**0.1**	**TN**

What happens if I set my threshold at 0.5?

True positive rate =

False alarm rate =

Precision =

Recall =

Prediction and thresholds

Beer	Grass	Rice	Flood	Prediction	Metric
100	3	3	**Yes**	**0.8**	**TP**
20	1	1	**Yes**	**0.3**	**FN**
80	3	2	**No**	**0.6**	**FA**
40	1	1	**No**	**0.2**	**TN**
40	1	1	**No**	**0.1**	**TN**

What happens if I set my threshold at 0.5?

True positive rate = ½

False alarm rate = ⅓

Precision = ½

Recall = ½

Prediction and thresholds

Beer	Grass	Rice	Flood	Prediction	Metric
100	3	3	**Yes**	**0.8**	
20	1	1	**Yes**	**0.3**	
80	3	2	**No**	**0.6**	
40	1	1	**No**	**0.2**	
40	1	1	**No**	**0.1**	

What happens if I set my threshold at 0.15?

Prediction and thresholds

Beer	Grass	Rice	Flood	Prediction	Metric
100	3	3	**Yes**	**0.8**	**TP**
20	1	1	**Yes**	**0.3**	**TP**
80	3	2	**No**	**0.6**	**FA**
40	1	1	**No**	**0.2**	**FA**
40	1	1	**No**	**0.1**	**TN**

What happens if I set my threshold at 0.15?

True positive rate = 1

False alarm rate = 2/3

Precision = 2/4

Recall = 2/2

Prediction and thresholds

Beer	Grass	Rice	Flood	Prediction	Metric
100	3	3	**Yes**	**0.8**	
20	1	1	**Yes**	**0.3**	
80	3	2	**No**	**0.6**	
40	1	1	**No**	**0.2**	
40	1	1	**No**	**0.1**	

What happens if I set my threshold at 0.5?

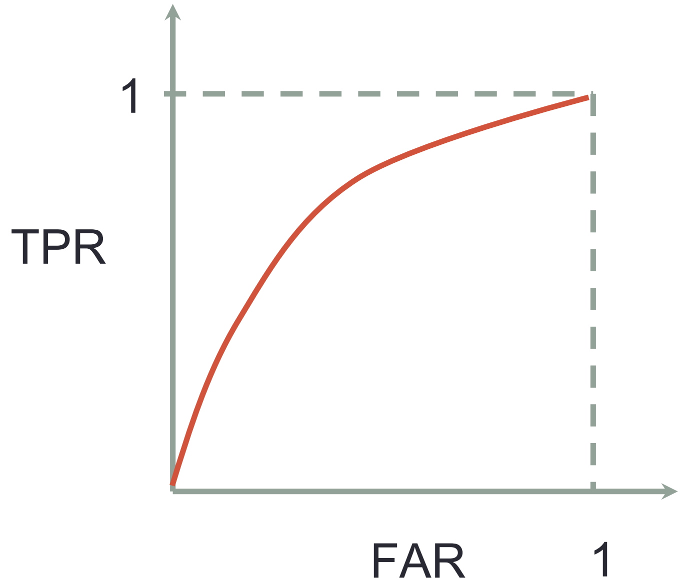
- True positive rate = ½
- False alarm rate = ⅓
- Precision = ½
- Recall = ½

What happens if I set my threshold at 0.15?

- True positive rate = 1
- False alarm rate = 2/3
- Precision = 2/4
- Recall = 2/2

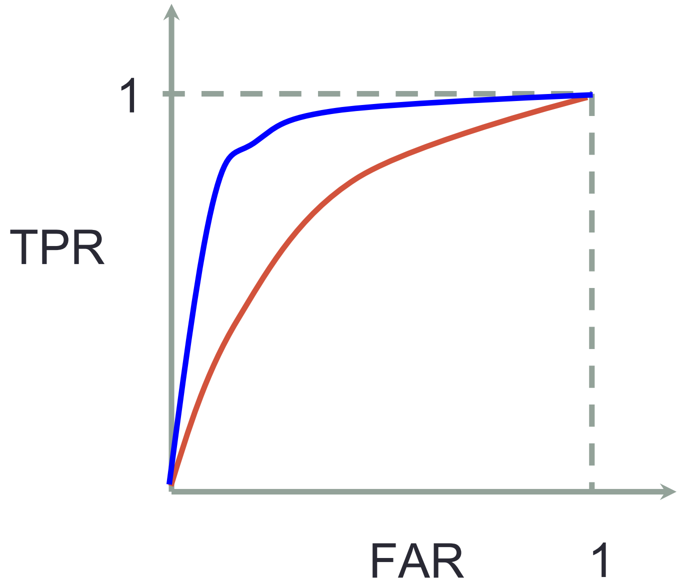
Receiver operating Characteristic (RoC) curve

- What if we change the threshold
- FA TP is a tradeoff
  This is why we need to think of the application when thinking of metrics.
- Plot FA rate and TP rate as threshold changes



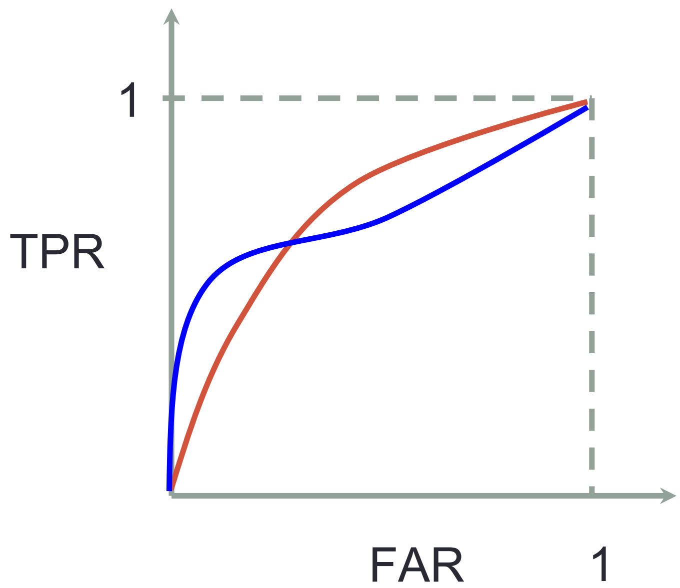
Comparing detectors

- Which is better?



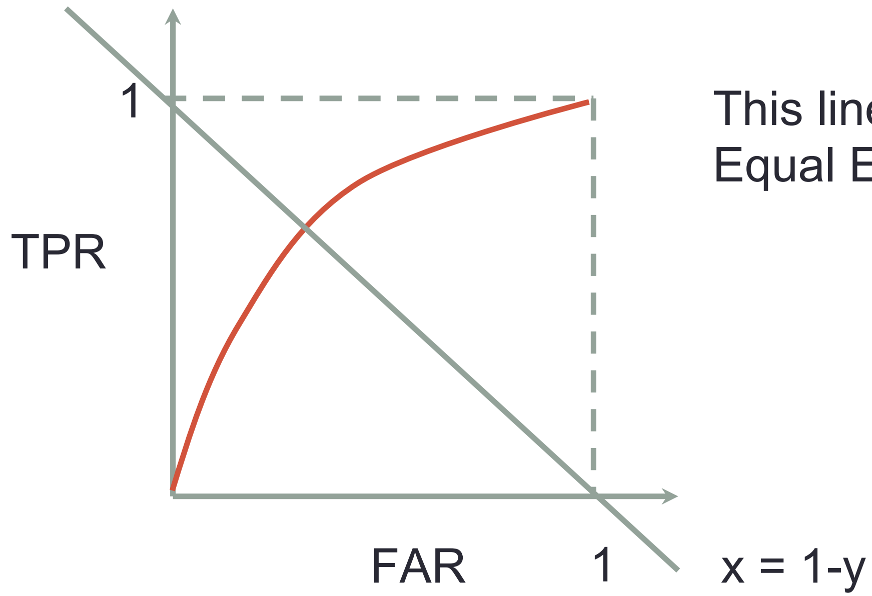
Comparing detectors

- Which is better?



Selecting the threshold

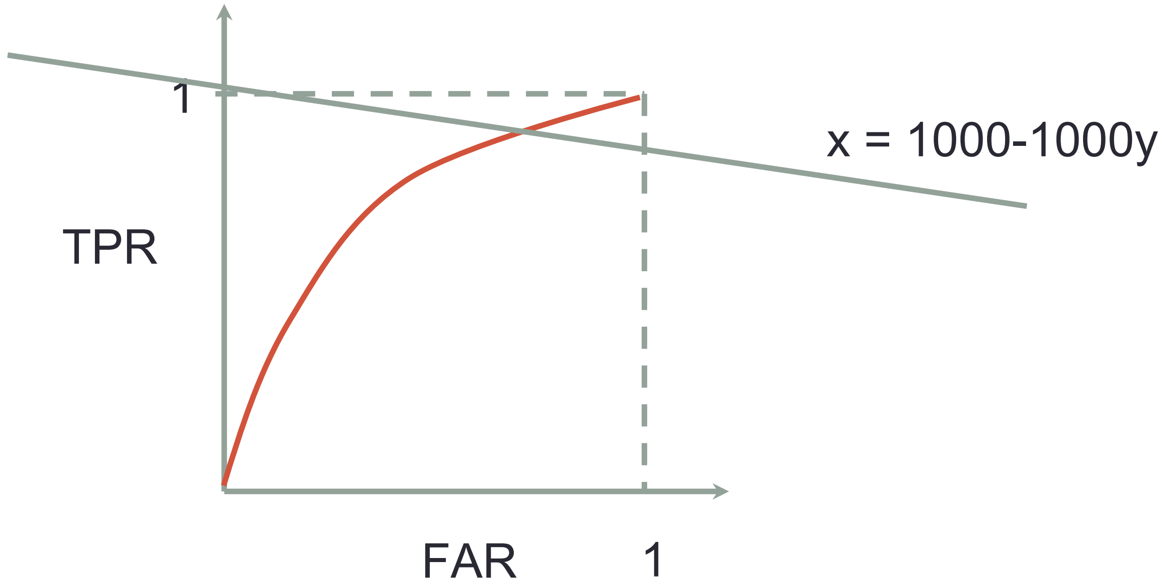
- Select based on the application
- Trade off between TP and FA. Know your application, know your users.
  - A miss is as bad as a false alarm   FAR = 1-TPR => x = 1-y



This line has a special name
Equal Error Rate (EER)

Selecting the threshold

- Select based on the application
- Trade off between TP and FA. Know your application, know your users. Is the application about safety?
  - A miss is 1000 times more costly than false alarm.
    - FAR = 1000(1-TPR) => x = 1000-1000y



Churn prediction

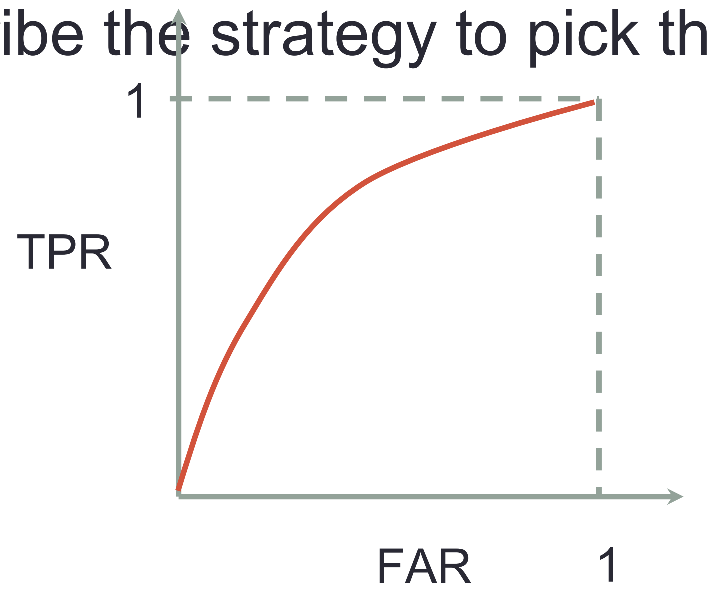
Predict whether a customer will stop subscription, so we can send a promotional ad.

Usual subscription fee 50

Cost of calling the customer 5

Promotional subscription fee 25

Describe the strategy to pick the threshold



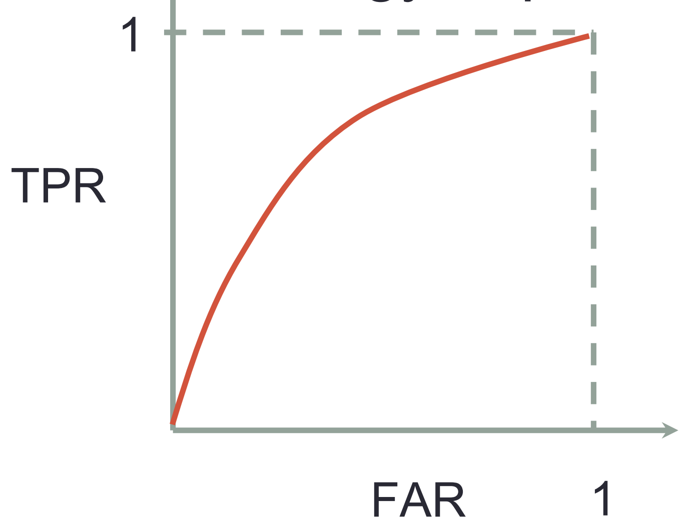
Churn prediction

Predict whether a customer will stop subscription, so we can send a promotional ad.

Usual subscription fee 50

Cost of calling the customer 5

- Promotional subscription fee 25
  - Describe the strategy to pick the threshold

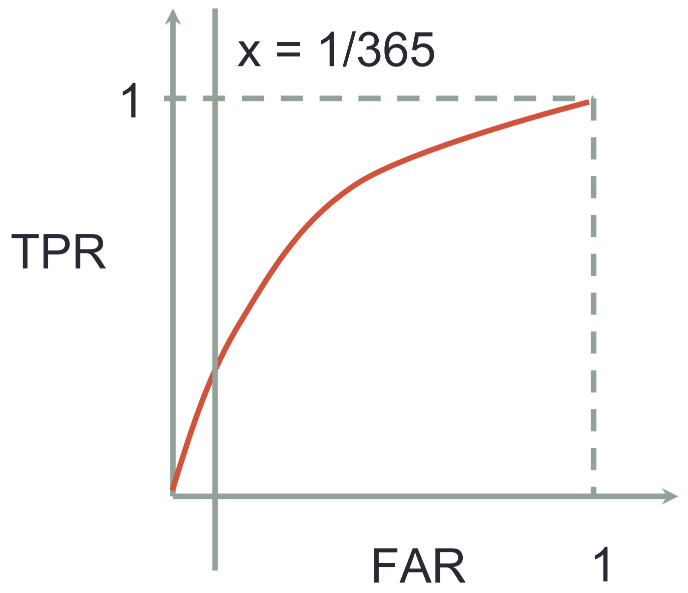


Cost of miss = 50
Cost of FA = 30

30FAR = 50(1-TPR)

Selecting the threshold

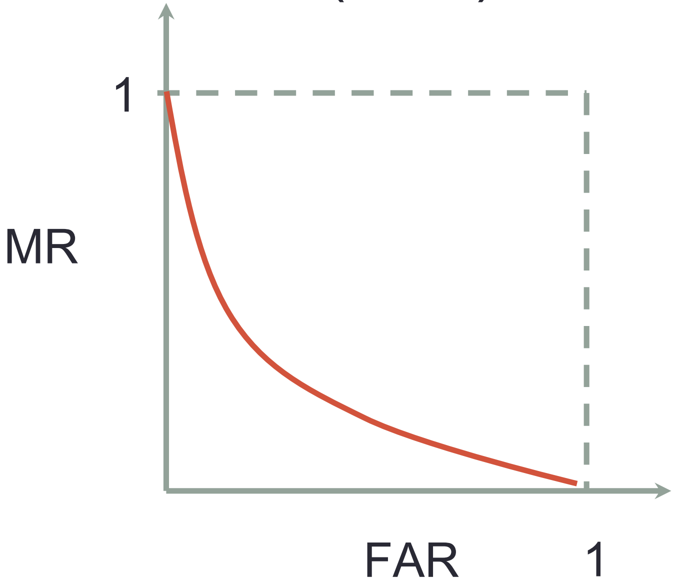
- Select based on the application
- Trade off between TP and FA.
  - Regulation or hard threshold
  - Cannot exceed 1 False alarm per year
    - If 1 decision is made everyday, FAR = 1/365



Notes about RoC

- Ways to compress RoC to just a number for easier comparison -- use with care!!
  - EER
  - Area under the curve
  - F score
- Other similar curve - Detection Error Tradeoff (DET) curve
  - Plot False alarm vs Miss rate
- Other similar curve

  PR curve (precision-recall curve)
- Can plot on log scale for clarity



Summary

- Train-validation-test
- Hyperparameter vs parameter
- RoC