Exam: 22.05.2019, 10:00-13:00, Wednesday
OBIKAS course evaluation
Project presentations: 31 May, Friday, 9:00-12:00, A2

Design and Analysis of Machine Learning Experiments

Emre Ugur, BM 33

emre.ugur@boun.edu.tr

http://www.cmpe.boun.edu.tr/~emre/courses/cmpe462

cmpe462@listeci.cmpe.boun.edu.tr

Quiz

Initialize all $Q(s, a)$ arbitrarily
For all episodes
    Initalize s
    Repeat
        Choose a using policy derived from Q, e.g., ϵ-greedy
        Take action a, observe r and s'
        Update $Q(s, a)$:
            $Q(s, a) \leftarrow Q(s, a) + \eta(r + \gamma \max_{a'} Q(s', a') - Q(s, a))$
        $s \leftarrow s'$
    Until s is terminal state

or a = π (s)

Figure 18.5 Q learning, which is an off-policy temporal difference algorithm.

Which action selection method you use to learn the optimal policy in the following problem? Why?

+10	+10	+10	+5	10	10	10	100
+10		+20		+50			
S	-10	-15	-20	-30	-40	-50	+1000 00000

Performance

- Aim of Machine Learning studies:
  - Assessment of the expected error of a learning algorithm: Is the error rate of 1-NN less than 2%?
  - Comparing the expected errors of two algorithms: Is k-NN more accurate than MLP ?
- Training error?
  - For comparison? More complex always..
- Validation set
  - One run?
    - Small sets, exceptions, outliers
    - Randomness.
- Use a distribution of validation errors.
  - Generate multiple learners, test on multiple validation sets.

Keep in mind

- Free lunch theorem: There is no one best learner for all problems.
- Split into training and validation is for selecting the best. After selecting best hyper-parameters, learning method..
- 3 subsets:
  - Training: optimize the parameters
  - Validation: optimize the method and its hyper-parameters
  - Test: report the error
- Not only error-rate base, what are the other metrics?
  - Risks that depend on loss function
  - Training time and space complexity
  - Testing time and space complexity
  - Interpretability
  - Easy programmability

Cross-Validation and Resampling Methods

- Repeatedly use the same data split differently. K training/validation set pairs: $\{\mathcal{T}_i, \mathcal{V}_i\}_{i=1}^{K}$
- Ensure stratification: class prior probabilities.
- K-fold cross-validation: K equal sized sets: $\mathcal{X}_i, i = 1, \ldots, K$

$$
\begin{aligned}
\mathcal{V}_1 &= \mathcal{X}_1 \quad & \mathcal{T}_1 &= \mathcal{X}_2 \cup \mathcal{X}_3 \cup \cdots \cup \mathcal{X}_K \\
\mathcal{V}_2 &= \mathcal{X}_2 \quad & \mathcal{T}_2 &= \mathcal{X}_1 \cup \mathcal{X}_3 \cup \cdots \cup \mathcal{X}_K \\
&\vdots & & \\
\mathcal{V}_K &= \mathcal{X}_K \quad & \mathcal{T}_K &= \mathcal{X}_1 \cup \mathcal{X}_2 \cup \cdots \cup \mathcal{X}_{K-1}
\end{aligned}
$$

- allow small validation sets, significant overlap between training sets.
- 5 x 2 cross-validation: same sized training & validation sets. Divide into two, then swap. Shuffle, divide into two, then swap...

Measuring classifier performance

- Four possible cases in two class problems:

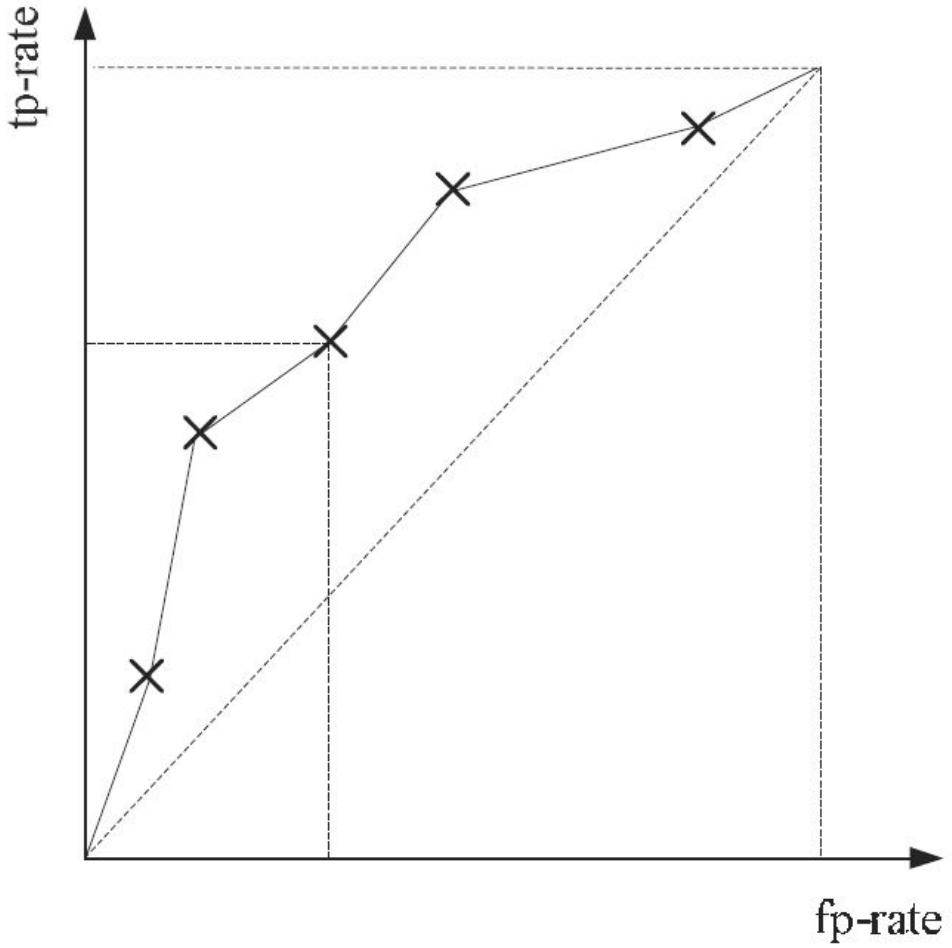
	Predicted class		
True Class	Positive	Negative	Total
Positive	tp : true positive	fn : false negative	p
Negative	fp : false positive	tn : true negative	n
Total	p'	n'	N

e.g. Voice authentication - high fp or fn?

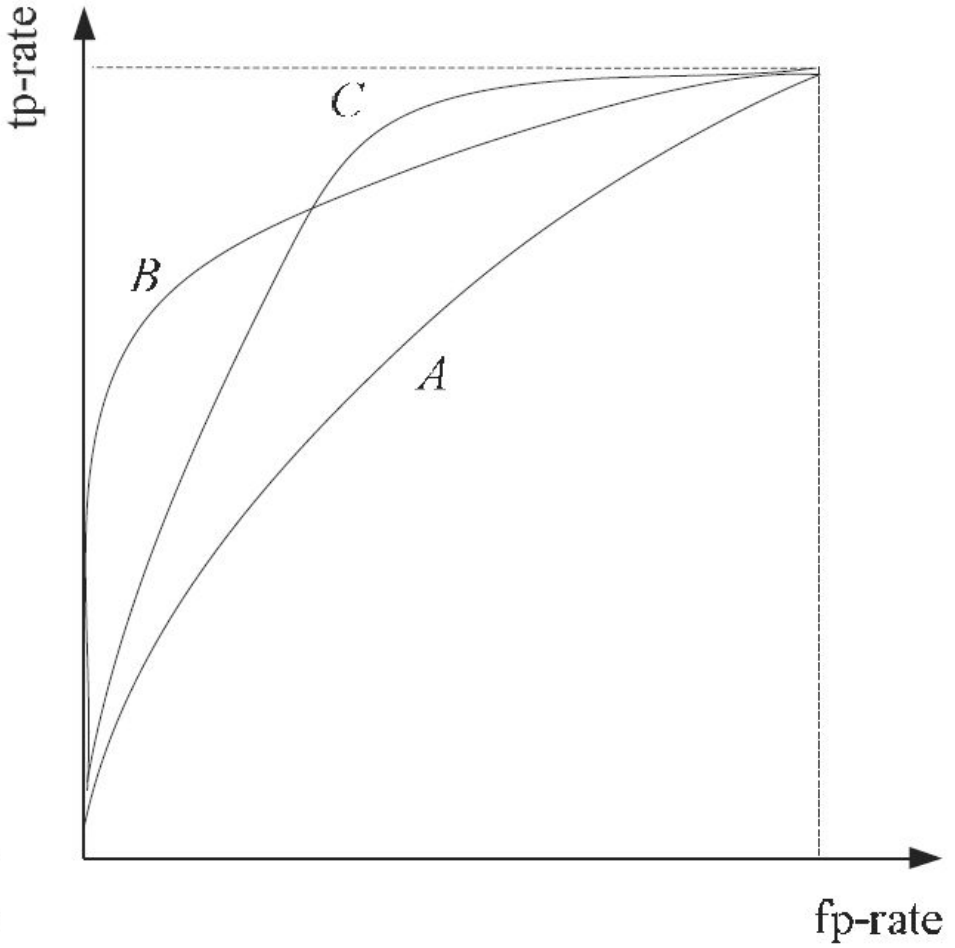
- Different measures in two class:

Name	Formula
error	$(fp + fn)/N$
accuracy	$(tp + tn)/N = 1 - \text{error}$
tp-rate	tp/p
fp-rate	fp/n
precision	tp/p'
recall	$tp/p = \text{tp-rate}$
sensitivity	$tp/p = \text{tp-rate}$
specificity	$tn/n = 1 - \text{fp-rate}$

Fine-tune classifier: plot Receiver Operating Characteristics (ROC) curve



(a) Example ROC curve



(b) Different ROC curves for different classifiers

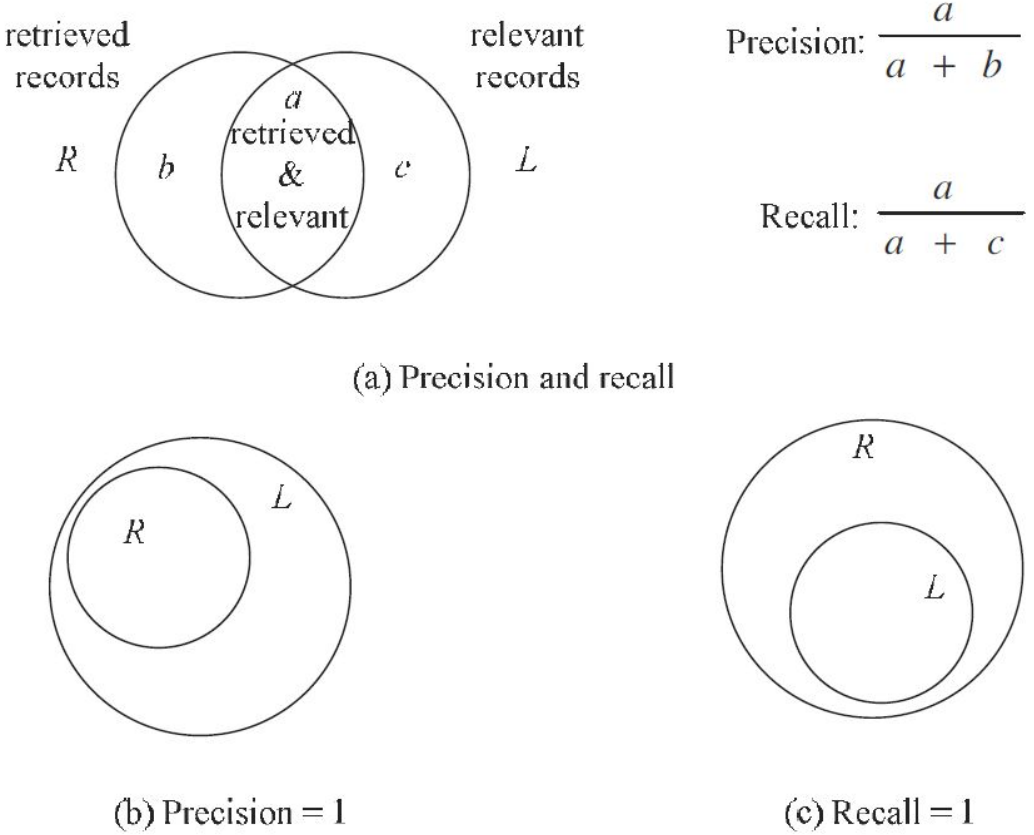
tp-rate	tp/p
fp-rate	fp/n

http://www.navan.name/roc/

Information retrieval

- Bring particular records:
  - True positives: Correctly retrieved relevant
  - False negatives: No retrieved but relevant
  - False positives: Retrieved irrelevant

Name	Formula
error	$(fp + fn)/N$
accuracy	$(tp + tn)/N = 1 - \text{error}$
tp-rate	tp/p
fp-rate	fp/n
precision	tp/p'
recall	$tp/p = \text{tp-rate}$
sensitivity	$tp/p = \text{tp-rate}$
specificity	$tn/n = 1 - \text{fp-rate}$



Measuring classifier performance

- Four possible cases in two class problems:

	Predicted class		
True Class	Positive	Negative	Total
Positive	tp : true positive	fn : false negative	p
Negative	fp : false positive	tn : true negative	n
Total	p'	n'	N

- Different measures in two class:

Name	Formula
error	$(fp + fn)/N$
accuracy	$(tp + tn)/N = 1 - \text{error}$
tp-rate	tp/p
fp-rate	fp/n
precision	tp/p'
recall	$tp/p = \text{tp-rate}$
sensitivity	$tp/p = \text{tp-rate}$
specificity	$tn/n = 1 - \text{fp-rate}$

Measuring classifier performance

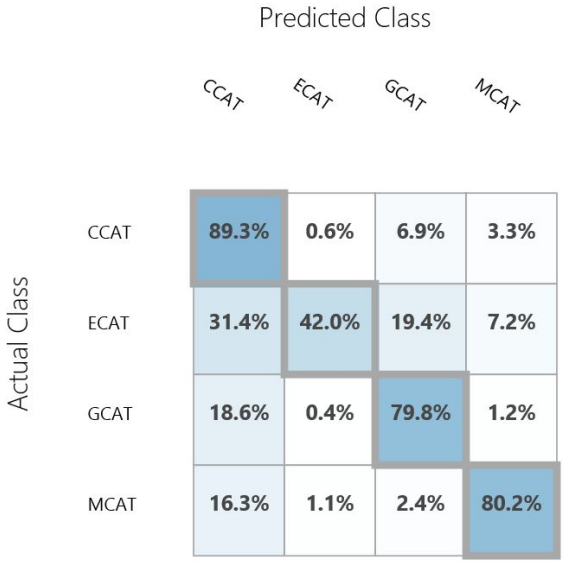
- Four possible cases in two class problems:

	Predicted class		
True Class	Positive	Negative	Total
Positive	tp : true positive	fn : false negative	p
Negative	fp : false positive	tn : true negative	n
Total	p'	n'	N

- Different measures in two class:

Name	Formula
error	$(fp + fn)/N$
accuracy	$(tp + tn)/N = 1 - \text{error}$
tp-rate	tp/p
fp-rate	fp/n
precision	tp/p'
recall	$tp/p = \text{tp-rate}$
sensitivity	$tp/p = \text{tp-rate}$
specificity	$tn/n = 1 - \text{fp-rate}$

- Multi-class problems:
  - class confusion matrix



Confusion matrix

With LBP features

	Anger (%)	Disgust (%)	Fear (%)	Joy (%)	Sadness (%)	Surprise (%)	Neutral (%)
Anger	**58.7**	5.5	0	0	26.7	0	9.1
Disgust	3.3	**85.0**	2.5	0	2.5	0	6.7
Fear	1.0	0	**61.7**	24.0	10.3	0	3.0
Joy	0	0	6.0	**90.4**	0	0	3.6
Sadness	4.9	0	0	0	**72.4**	1.7	21.0
Surprise	0	0	1.3	0	2.7	**92.4**	3.6
Neutral	2.0	0.8	0.4	0.8	25.7	0	**70.3**