The Introduction To Artificial Intelligence

Yuni Zeng yunizeng@zstu.edu.cn
2022-2023-1

The Introduction to Artificial Intelligence

- Part I Brief Introduction to AI & Different AI tribes
- Part II Knowledge Representation & Reasoning
- Part III AI GAMES and Searching
- Part IV Model Evaluation and Selection

Model Evaluation and Selection

- 1.1 Empirical Error and Overfitting
- 1.2 Evaluation Methods
- 1.3 Performance Measure
- 1.4 Summary

Model Evaluation and Selection

Solve two problems:
(1) How to make a model convincible?
(2) How to evaluate a model?

Model Evaluation and Selection

- 1.1 Empirical Error and Overfitting
- 1.2 Evaluation Methods
- 1.3 Performance Measure
- 1.4 Summary

1.1 Empirical Error and Overfitting

Definitions

Usually, if m samples totally, a model predict a samples incorrectly:

Error Rate: a/m, the proportion of wrong predictions;

Accuracy: 1- a/m, the proportion of right predictions.

Generally:

Error: the difference between the output of the model and the ground truth (real label).

Training Error/ Empirical Error: the error on training dataset.

Generalization Error: the error on new data.

1.1 Empirical Error and Overfitting

Definitions

- Error Rate, Accuracy, Training error, Generalization error

- A best model:

  On training dataset: 0% Training error, 100% Accuracy,

  On new samples: 0% Generalization error, 100% Accuracy

- But:

0% Training error

0% Generalization error

1.1 Empirical Error and Overfitting

Overfitting and Underfitting

- Underfitting: A statistical model or a machine learning algorithm is said to have underfitting when it cannot capture the underlying trend of the data, i.e., it only performs well on training data but performs poorly on testing data.

- In a nutshell, Underfitting refers to a model that can neither performs well on the training data nor generalize to new data.

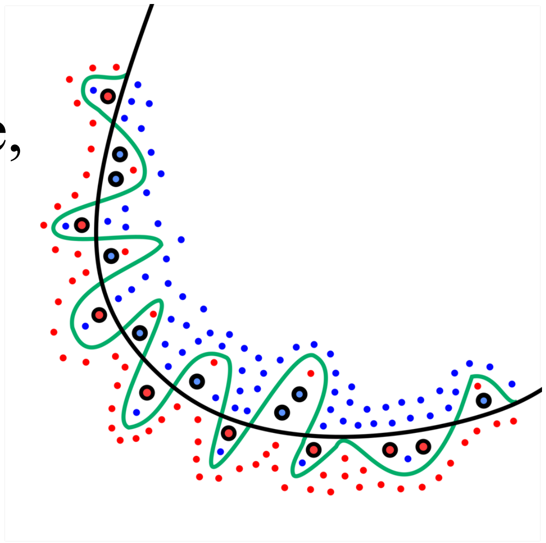
- Reasons for Underfitting:
    - High bias and low variance
    - The size of the training dataset used is not enough.
    - The model is too simple.
    - Training data is not cleaned and also contains noise in it.

1.1 Empirical Error and Overfitting

Overfitting and Underfitting

- Overfitting: In mathematical modeling, **overfitting** is "the production of an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit to additional data or predict future observations reliably".

- An overfitted model is a mathematical model that contains more parameters than can be justified by the data. In a mathematical sense, these parameters represent the degree of a polynomial. The essence of overfitting is to have unknowingly extracted some of the residual variation (i.e., the noise) as if that variation represented underlying model structure.



1.1 Empirical Error and Overfitting

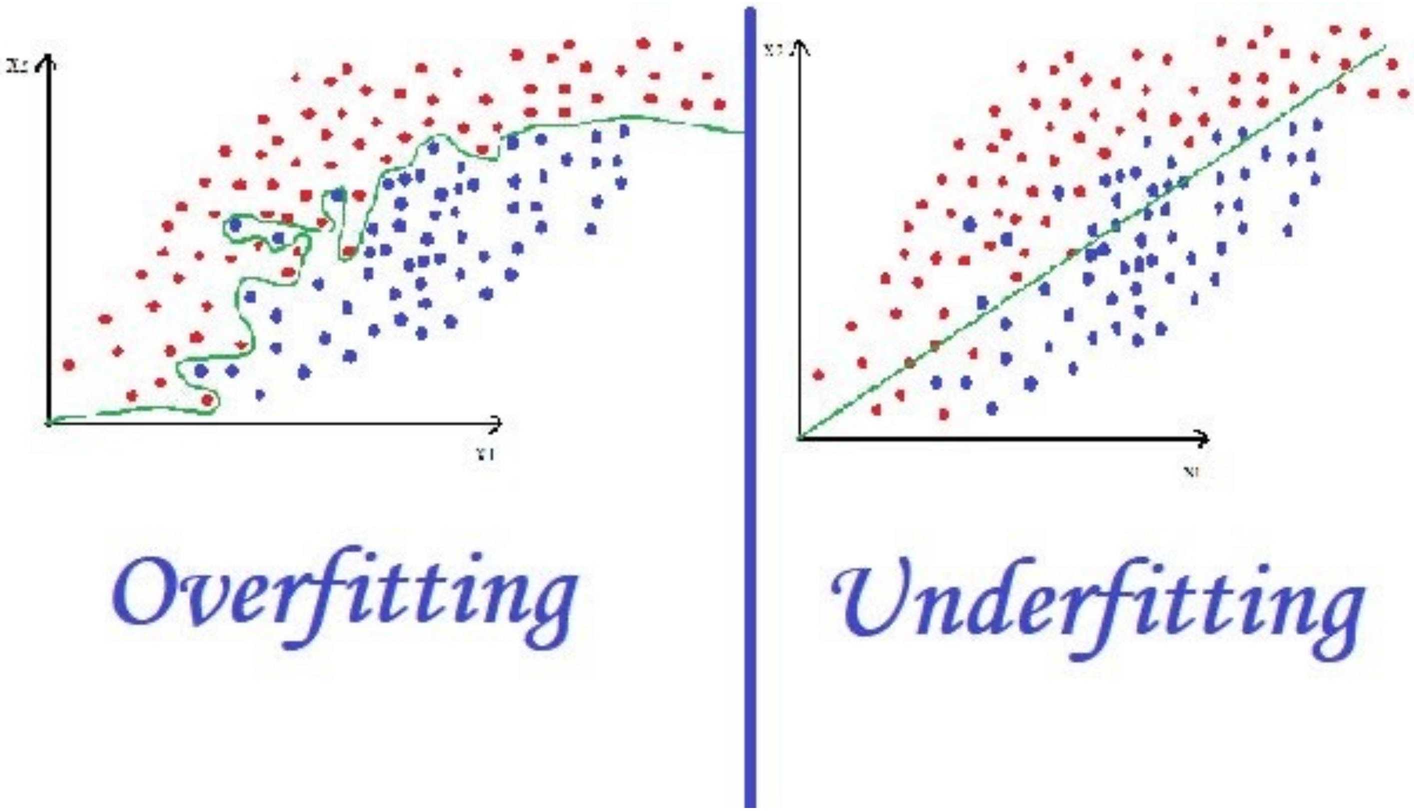
Overfitting and Underfitting

- In a nutshell, Overfitting is a problem where the evaluation of machine learning algorithms on training data is different from unseen data.

- Reasons for Overfitting are as follows:
  - High variance and low bias
  - The model is too complex
  - The size of the training data

1.1 Empirical Error and Overfitting

- Overfitting and Underfitting

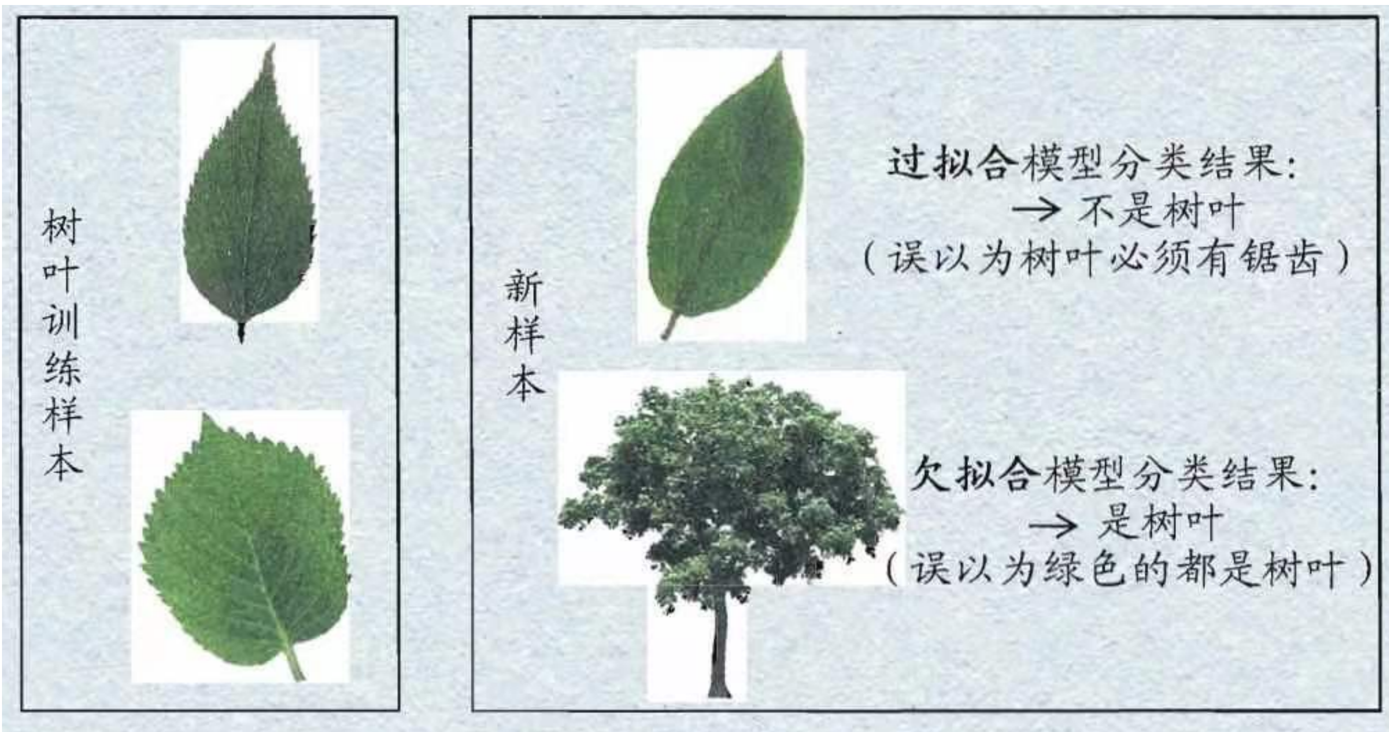


1.1 Empirical Error and Overfitting

Overfitting and Underfitting

Training samples

News samples



Overfitting

Underfitting

- Chapter 2 Model Evaluation and Selection, Machine Learning, Zhihua Zhou

1.1 Empirical Error and Overfitting

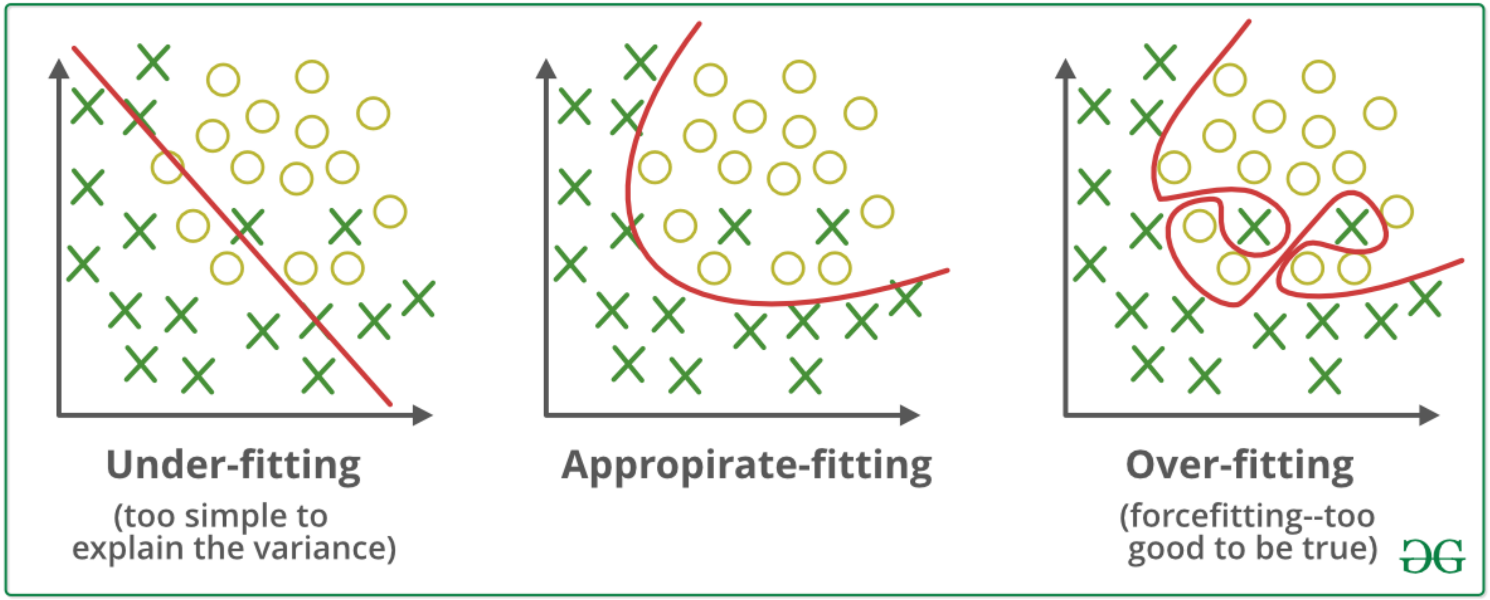
Overfitting and Underfitting

- **Techniques to reduce underfitting:**
  - Increase model complexity
  - Increase the number of features, performing feature engineering
  - Remove noise from the data.
  - Increase the number of epochs or increase the duration of training to get better results.
- **Techniques to reduce overfitting:**
  - Increase training data.
  - Reduce model complexity.
  - Early stopping during the training phase (have an eye over the loss over the training period as soon as loss begins to increase stop training).
  - Ridge Regularization and Lasso Regularization
  - Use dropout for neural networks to tackle overfitting.

1.1 Empirical Error and Overfitting

Overfitting and Underfitting

- Overfitting: Good performance on the training data, poor generalization to other data.
- Underfitting: Poor performance on the training data and poor generalization to other data.



Model Evaluation and Selection

We already know:
What kind of model do we need?
→ Low training error, low generalization error, high Accuracy;
→ However, many methods could be utilized for one problem with different parameters.
→ how to select a model?

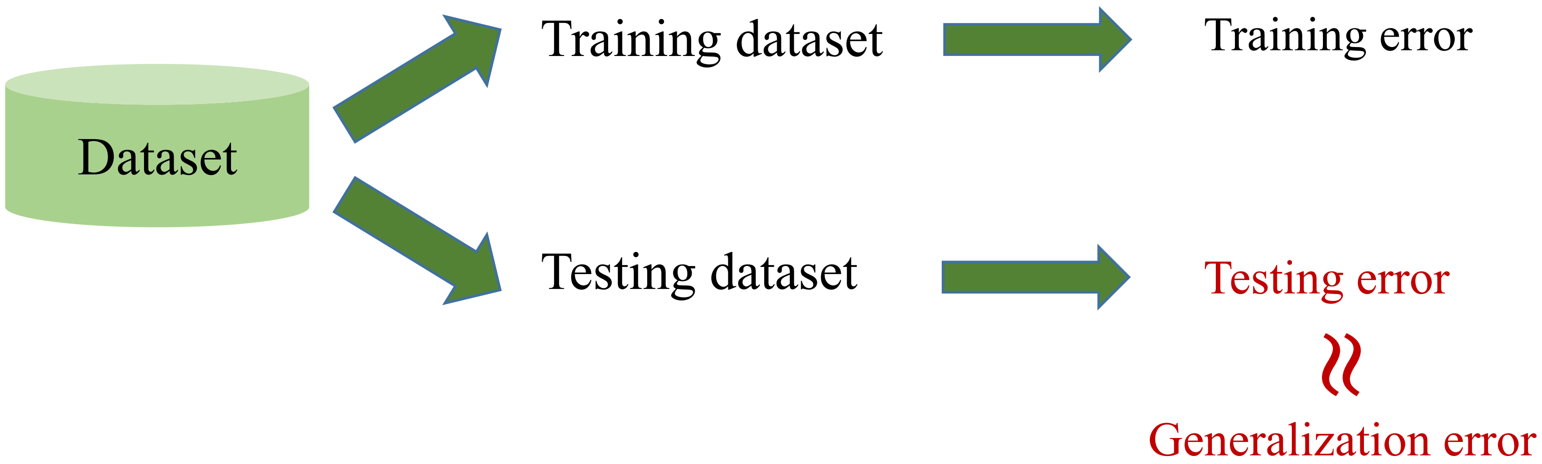
Model Evaluation and Selection

- 1.1 Empirical Error and Overfitting
- 1.2 Evaluation Methods
- 1.3 Performance Measure
- 1.4 Summary

1.2 Evaluation Methods

Evaluation Methods

- A model with low training error, low generalization error, high accuracy;
- How to compute generalization error?



How to divide training dataset and testing dataset?

1.2 Evaluation Methods

Evaluation Methods

- For example, m samples:

$$D = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), \cdots, (x_m, y_m)\}$$

Training dataset: S Testing dataset: T

- $S \cap T = \emptyset$
- $S \cup T = D$

How to divide training dataset and testing dataset?

1.2 Evaluation Methods

Hold-out Method (留出法)

- Set a proportion r, like $r = 0.3$
- By sampling methods, make

$$T = r * D, S = (1 - r) * D$$

- Sampling methods:
  - Random sampling
  - Stratified sampling: keep the proportion rate of samples;

    For example, 500 positive samples, 500 negative samples in D and $r = 0.3$:
    S: 350 positive samples; 350 negative samples
    T: 150 positive samples; 150 negative samples

- Difficulty: r,

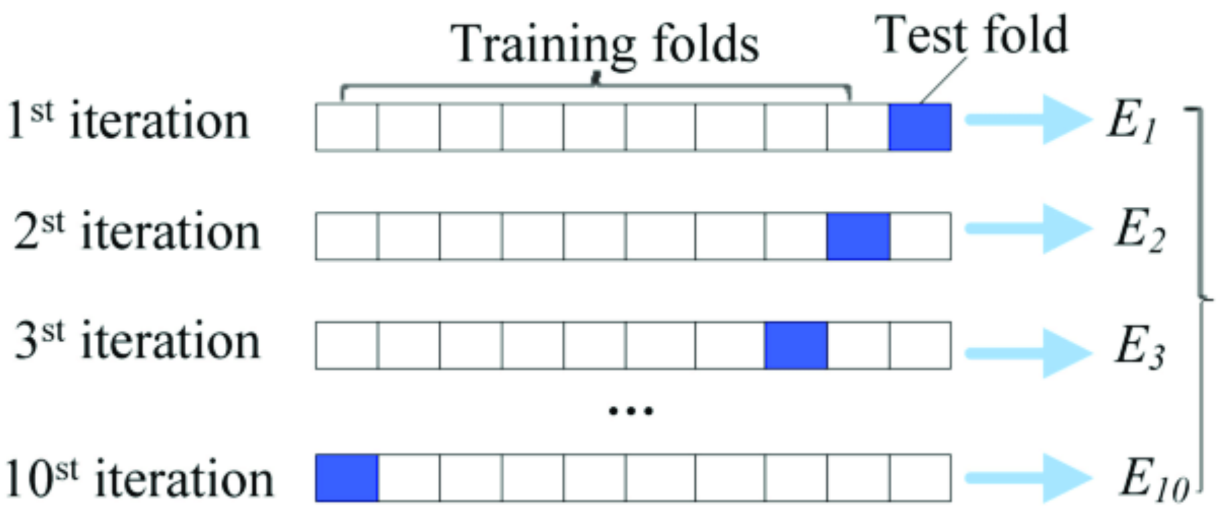
1.2 Evaluation Methods

Cross Validation (交叉验证法)

- Divided D dataset to k sub-dataset:

$$D = D_1 \cup D_2 \cup \cdots D_k,\ D_i \cap D_j = \emptyset\ (i \neq j)$$

- Keep same distribution of each sub-dataset
- K-times test: (k-1) sub-dataset as S, 1 sub-dataset as T
- Average K-times test error as final results.

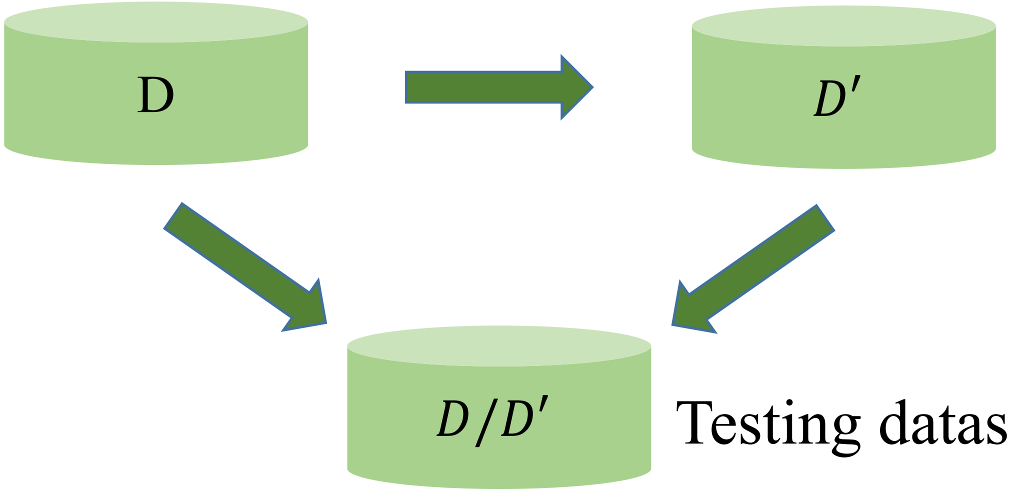


$$E = \frac{1}{k}\sum_{i=1}^{k} E_i$$

1.2 Evaluation Methods

Bootstrapping (自助法)

- Based on Bootstrapping Sampling
  - Randomly select 1 sample from D and copy it to D';
  - Repeat m times
  - Obviously, some of the samples in D will be repeated in D', and some will not.
  - Suitable for small datasets!



Training dataset: S

Testing dataset: T

Model Evaluation and Selection

- 1.1 Empirical Error and Overfitting
- 1.2 Evaluation Methods
- 1.3 Performance Measure
- 1.4 Summary

1.3 Performance Measure

Performance Measure

- For dataset D, with x_i as input, y_i as true label,

$$D = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), \cdots, (x_m, y_m)\}$$

- The predicted output of a model (f),

$$outputs = \{y_1^*, y_2^*, y_3^*, \cdots, y_m^*\}$$

How to measure the predictions of different models?

1.3 Performance Measure

- Two different tasks
  - For regression tasks: Mean Squared Error

$$E(f, D) = \frac{1}{m}\sum_{i=1}^{m}(y_i^* - y_i)^2$$

  - The classification task: Error Rate, Accuracy

$$E(f, D) = \frac{1}{m}\sum_{i=1}^{m}\prod(y_i^* \neq y_i)$$

$$Acc(f, D) = \frac{1}{m}\sum_{i=1}^{m}\prod(y_i^* = y_i) = 1 - E(f, D)$$

1.3 Performance Measure

Confusion matrix

- For binary classification tasks

Sensitivity
Specificity

Decision /action	True state/class	
	Positive	Negative
Positive		
Negative		

Type-I Error
Type-II Error

Correct classification
TP: the number of samples belonging to positive decided positive
TN: the number of samples belonging to negative decided negative

Misclassification
FP: the number of samples belonging to negative decided positive incorrectly. (False Alarm)
FN: the number of samples belonging to positive decided negative incorrectly.(Missed Detection)

1.3 Performance Measure

- Sensitivity (TP rate)
  - $S_n = \frac{TP}{TP+FN}$
- Specificity (TN rate)
  - $S_p = \frac{TN}{TN+FP}$
- FP rate (Type-I Error)
  - FP rate = $\frac{FP}{FP+TN}$
- FN rate (Type-II Error)
  - FN rate = $\frac{FN}{FN+TP}$

Decision/ action	True state/class	
	Positive	Negative
Positive	TP	FP
Negative	FN	TN

$TP + FP + TN + FN = Total\ number\ of\ samples\ in\ dataset$

- Accuracy

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

- Precision

$$precision = \frac{TP}{TP + FP}$$

1.3 Performance Measure

- Sensitivity (TP rate)
  - $S_n = \frac{TP}{TP+FN}$
- Specificity (TN rate)
  - $S_p = \frac{TN}{TN+FP}$
- FP rate (Type-I Error)
  - FP rate = $\frac{FP}{FP+TN}$
- FN rate (Type-II Error)
  - FN rate = $\frac{FN}{FN+TP}$
- Accuracy

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

- Precision

$$precision = \frac{TP}{TP + FP}$$

Decision/ action	True state/class	
	Positive	Negative
Positive	TP	FP
Negative	FN	TN

$TP + FP + TN + FN =$
$Total\ number\ of\ samples\ in\ dataset$

1.3 Performance Measure

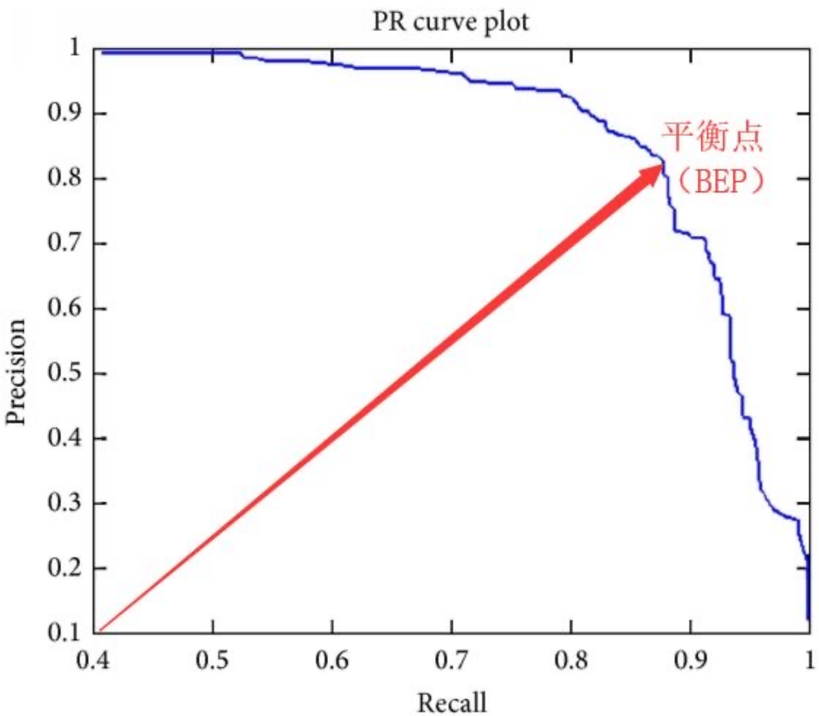
- Confusion matrix from Wiki

Sources: [21][22][23][24][25][26][27][28][29] view · talk · edit

		Predicted condition			
	Total population $= P + N$	**Positive (PP)**	**Negative (PN)**	Informedness, bookmaker informedness (BM) $= TPR + TNR - 1$	Prevalence threshold (PT) $= \frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	**Positive (P)**	**True positive (TP), hit**	**False negative (FN), type II error, miss, underestimation**	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate $= \frac{FN}{P} = 1 - TPR$
	Negative (N)	**False positive (FP), type I error, false alarm, overestimation**	**True negative (TN), correct rejection**	False positive rate (FPR), probability of false alarm, fall-out $= \frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$
	Prevalence $= \frac{P}{P + N}$	Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$	False omission rate (FOR) $= \frac{FN}{PN} = 1 - NPV$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$
	Accuracy (ACC) $= \frac{TP + TN}{P + N}$	False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$	Negative predictive value (NPV) $= \frac{TN}{PN} = 1 - FOR$	Markedness (MK), deltaP (Δp) $= PPV + NPV - 1$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$
	Balanced accuracy (BA) $= \frac{TPR + TNR}{2}$	F_1 score $= \frac{2PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$	Fowlkes–Mallows index (FM) $= \sqrt{PPV \times TPR}$	Matthews correlation coefficient (MCC) $= \sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$	Threat score (TS), critical success index (CSI), Jaccard index $= \frac{TP}{TP + FN + FP}$

1.3 Performance Measure

- P-R Curve: Precision- Recall
  - A PR curve is simply a graph with Precision values on the y-axis and Recall （Sensitivity） values on the x-axis.



- The point is called "Break-Even Point, BEP", when precision= recall.
- If the BEP value of model A is bigger than it of model B, we can say model A is better than model B based on BEP.

1.3 Performance Measure

- Sensitivity (Recall, R)
  - $S_n = \frac{TP}{TP+FN}$
- Precision (P)
  - $precision = \frac{TP}{TP+FP}$

Decision/ action	True state/class	
	Positive	Negative
Positive	TP	FP
Negative	FN	TN

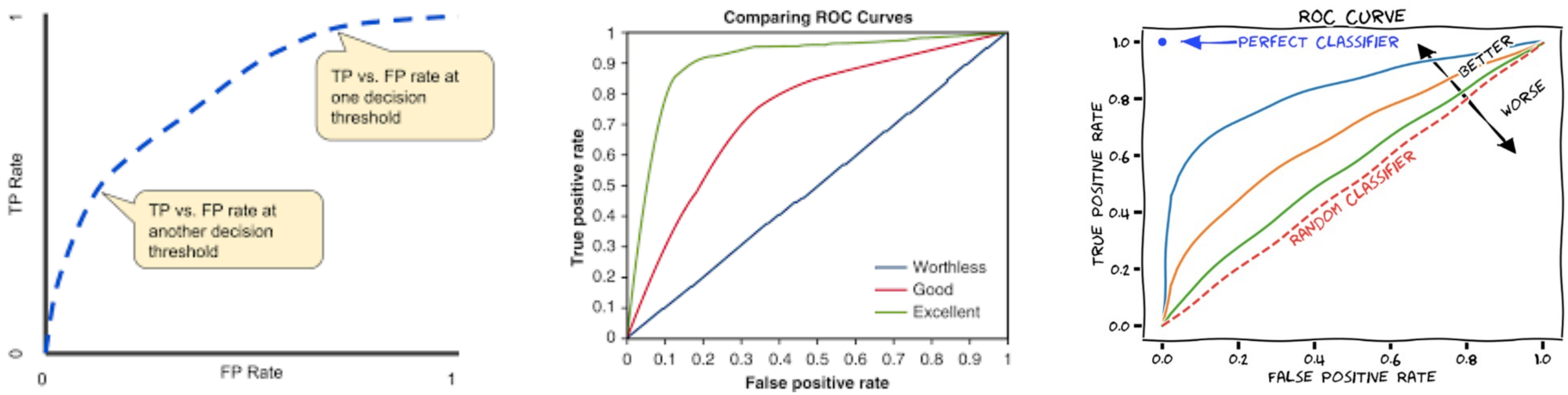
$TP + FP + TN + FN =$
$Total\ number\ of\ samples\ in\ dataset$

- F1

$$F1 = \frac{2 \times P \times R}{P + R}$$

1.3 Performance Measure

ROC Curve (Receiver Operating Characteristic)



- An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds.

2.3 Type-I Error Probability & Type-II Error Probability

- ROC Curve (Receiver Operating Characteristic)
  - For a binary classification,
    - 5 positive samples, and prediction probability: (0.9,0.8,0.5,0.4,0.3)
    - 5 negative samples: （0.7,0.6,0.2,0.1,0.01)
    - Ranking:(0.9,0.8,0.7,0.6,0.5,0.4,0.3,0.2,0.1,0.01)

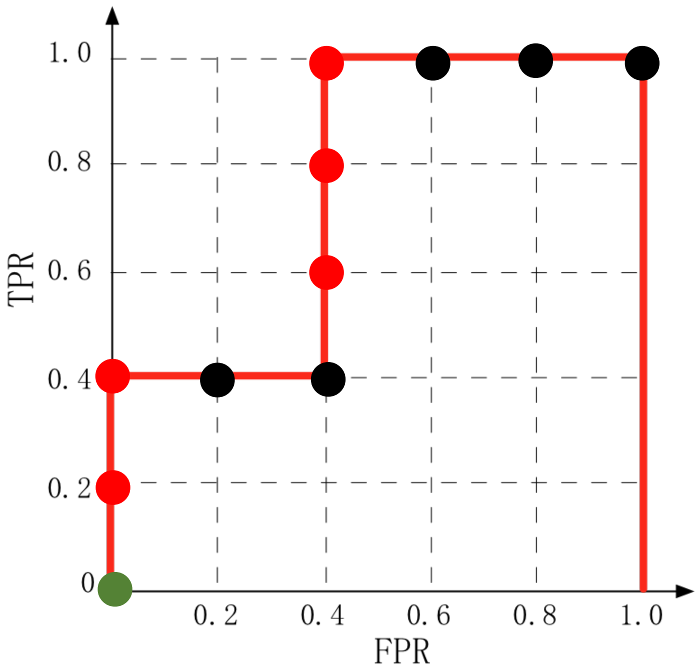
Thresholds	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.01
TPR = TP/(TP+FN)	0.2	0.4	0.4	0.4	0.6	0.8	1.0	1.0	1.0	1.0
FPR = FP/(FP+TN)	0	0	0.2	0.4	0.4	0.4	0.4	0.6	0.8	1.0

- FN: number of true positive samples; TP: number of true positive samples
- TN: number of true negative samples; FP: number of false positive samples

2.3 Type-I Error Probability & Type-II Error Probability

ROC Curve (Receiver Operating Characteristic)

Thresholds	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.01
TPR	0.2	0.4	0.4	0.4	0.6	0.8	1.0	1.0	1.0	1.0
FPR	0	0	0.2	0.4	0.4	0.4	0.4	0.6	0.8	1.0



- Area Under Curve: AUC
- AUC:

$$AUC = \frac{1}{2}\sum_{i=1}^{m-1}(x_{i+1} - x_i)\cdot(y_i + y_{i+1})$$

- AUC =1; perfect!
- 0.5<AUC<1, better than randomly classification;
- AUC = 0.5, same as randomly classification;

Test

- ROC Curve (Receiver Operating Characteristic)

样本编号	真实标签	模型输出概率	样本编号	真实标签	模型输出概率
1	p	0.9	11	p	0.4
2	p	0.8	12	n	0.39
3	n	0.7	13	p	0.38
4	p	0.6	14	n	0.37
5	p	0.55	15	n	0.36
6	p	0.54	16	n	0.35
7	n	0.53	17	p	0.34
8	n	0.52	18	n	0.33
9	p	0.51	19	p	0.30
10	n	0.505	20	n	0.10

- p : positive sample, n: negative sample

Summary

- 1.1 Empirical Error and Overfitting
- 1.2 Evaluation Methods
- 1.3 Performance Measure
- 1.4 Summary

Model Evaluation and Selection

Solve two problems:
(1) How to make a model convincible?
(2) How to evaluate a model?

Summary

- How to make a model convincible?
  - Error, Training error, Generalization error
  - Overfitting and Underfitting
  - Evaluation Methods: Hold-out method, Cross Validation, Bootstrapping
- How to evaluate a model?
  - Measure metrics: ACC, Recall, F1,AUC…