BOISE STATE UNIVERSITY

CS 533
INTRO TO DATA SCIENCE

Instructor: Jun Zhuang

Source: https://slideplayer.com/slide/15849901/

MODEL EVALUATION

INTRODUCTION

What Is Model Evaluation?

- A systematic assessment of how well a model performs on a given task, such as predictions or classifications.

- Can help us measure the quality of our models and make decisions in real-world scenarios.

Image source: generated by DALLE 2.

Why Is Model Evaluation Important?

Assessing model performance can help us:

- Fine-tune the models for improving the performance.
- Compare the model with different baseline models.
- Understand the model performance in unseen data.
- Select the suitable model for the specific tasks.

MODEL EVALUATION

MODEL TRAINING

Training Process

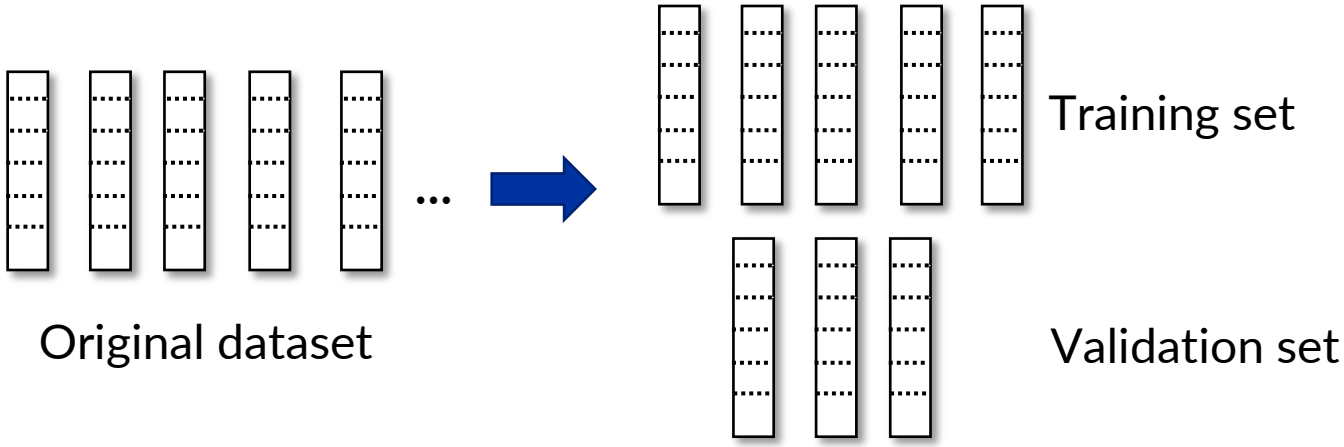
We train the model to learn the patterns in the train data and generalize the model to unseen data.

During this process, we typically:

- Feed preprocessed train data to the model.
- Optimize the model parameters via some optimization algorithms, such as gradient descent.
- Tune the model hyper-parameters based on the validation data.
- Store the trained model parameters (weights) or latent embedding for further use if necessary.

Model Validation

- Question: Since we hope to choose a model with smallest generalization error, how can we achieve our goal if we only have training samples?



Make sure that samples in the validation set do not appear in the training set!

Model Validation

- Question: Since we hope to choose a model with smallest generalization error, how can we achieve our goal if we only have training samples?

  Answer: We can construct a validation set from the given dataset and use the validation error as the approximation of a model's generalization error.

Make sure that samples in the validation set do not appear in the training set!

Hyper-parameter Tuning

- Different hyper-parameters result in different performance.
- During training, we use a validation set to evaluate the performance and then tune the hyper-parameters.
- In practice, we tune the hyper-parameters by choosing a range of numbers.
- In some case (we may train the model using a subset first), after the model has been learnt and hyper-parameters has been fixed, the model should be retrained using the full training set.

Question

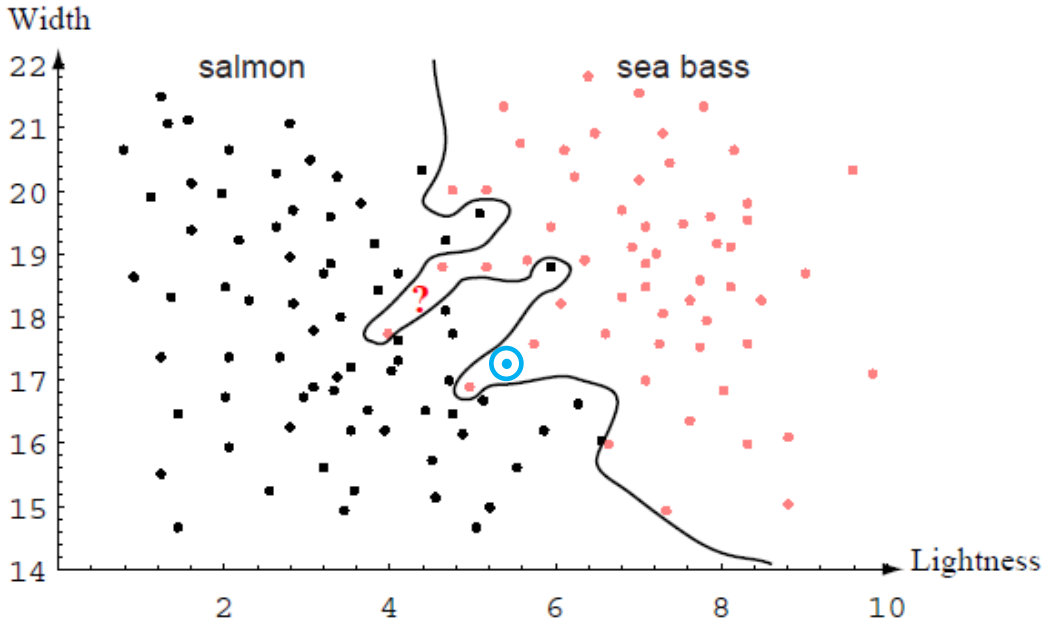
- What methods have you used to tune hyperparameters?

Training Error

- Error rate:
  - $E = \frac{\#\ of\ misclassified\ samples\ (a)}{\#\ of\ all\ samples\ (m)} = \frac{a}{m}$
- Accuracy:
  - $acc = 1 - \frac{a}{m} = 1 - E$
- Error is the difference between the output of a learner and the ground-truth of samples
- Training error, or empirical error, is the error of a learner on a training set

Generalization Error

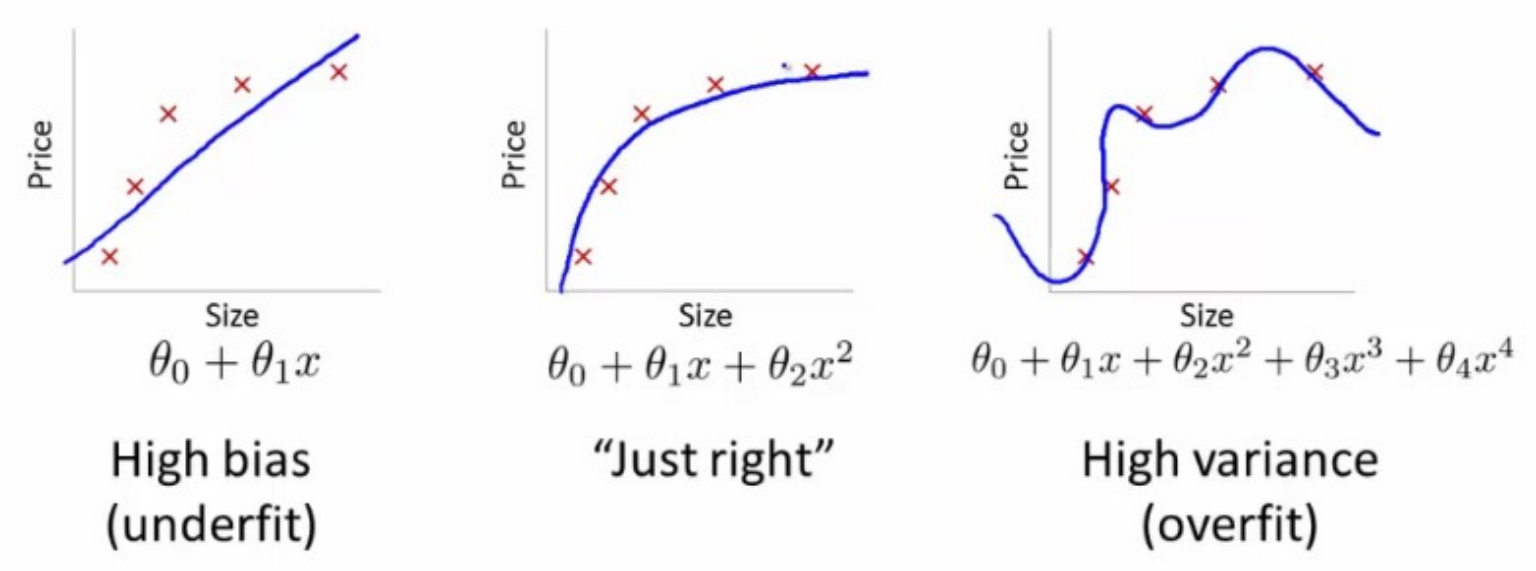
- What is a model's generalization ability?



- Generalization error is the error of a learner tested on a new dataset

Overfitting

- What kind of learners are good?
- Overfitting vs Underfitting.



Group Discussion

- Explain the bias and variance in your own words.
- As the model complexity increases, how will the bias and variance change (increase or decrease)?
- Explain the underfitting and overfitting in your own words.
- How to overcome the underfitting?
- How to avoid the overfitting?

MODEL EVALUATION

MODEL EVALUATION METHODS

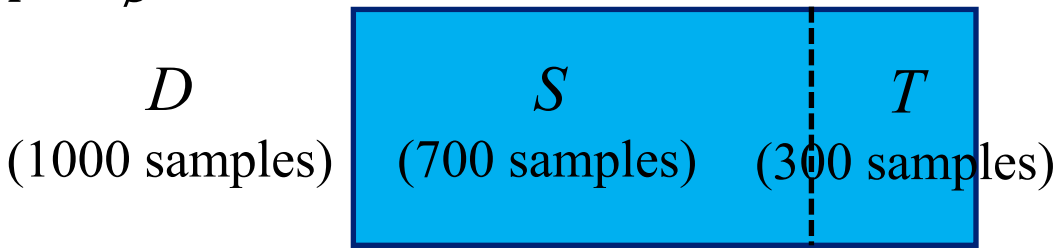
Model Evaluation Methods

Methods to construct training set and testing set from the original dataset:

- Hold-out
- Cross-validation
- Bootstrapping

Hold-out

- Given a dataset $D=\{(x_1, y_1), (x_2, y_2),\ldots,(x_m, y_m)\}$
- D is divided into two mutually exclusive sets: a training set S, and a testing set T, i.e. $D = S \cup T, S \cap T = \emptyset$



- Suppose there are 90 samples are misclassified.
- The error rate
  - $E = \frac{90}{300} * 100\% = 30\%$
- The accuracy $acc = 1 - 30\% = 70\%$.

Hold-out

- Noted that samples in the training set and testing set should obey the same distribution as samples distributed in the original dataset
- When constructing S and T, proportions of samples of difference classes in the training set, testing set, and the original dataset should be the same. (minimum requirement)
- Such kind of sampling method called **stratified sampling**.
- For example: If D has 500 positive samples and 500 negative samples, S should contain 350 pos samples and 350 neg samples, and T should contain 150 pos samples and 150 neg samples

Hold-out

- It's far from enough to fix the proportions of different classes in different sets.
- Different cut of D produces different training set and testing set.
- To evaluate the performance of a learner, we should randomly cut D for example 100 time and test the learner on 100 different testing sets (the size of 100 testing sets should be the same).
- The performance of a learner is the average performance on 100 testing set.

Hold-out

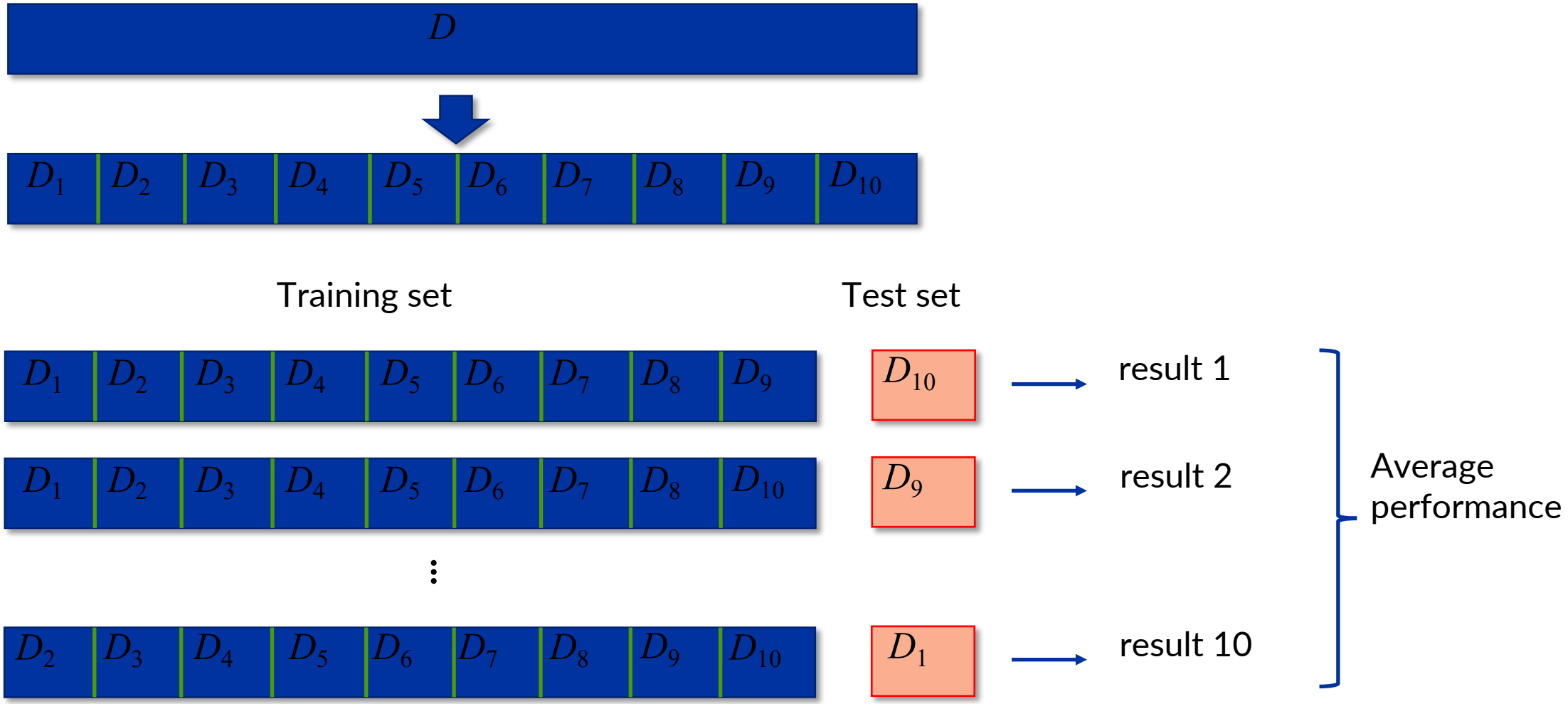
- Disadvantage
  - If S is too big, the learned model will be more close to the model trained using D. But because T is too small, the evaluation results may be not so accurate.
  - If S is small, the learned model may not be fully trained. Then the fidelity of evaluation results will be low.
- Usually 2/3~4/5 samples are used in the training process, and the left ones are used to test.

Cross-validation

- In cross-validation, D is divided into k mutually exclusive subset, i.e. $D = D_1 \cup D_2 \cup \cdots \cup D_k, D_i \cap D_j = \emptyset \ (i \neq j)$. D_i and D_j have the same distribution.

- k-fold cross-validation: The dataset is split into k subsets. The model is trained using the (k-1) subsets and evaluated using the k-th subset for each time. Repeat this process k times and report the average performance.

Cross-validation



Cross-validation

- Many ways to divide D into k subsets.
- Randomly divide D into k subsets and repeat the process p time. The final evaluation score is the average score of p times k-fold cross-validation.
- Special case of cross-validation, Leave-One-Out (LOO): Use one subset for evaluation.

Bootstrapping

- Resample instances from the existing dataset.
- Training set D -> D' (pick a sample from D m times)
- The probability of a sample a **not** picked is $(1\text{-}1/m)^m$
  - $\lim_{m\to\infty}\left(1-\frac{1}{m}\right)^m=\frac{1}{e}\approx 0.368$
    - D': training set
    - $D \setminus D'$: validation set.

Group Discussion

- What's the disadvantage of Leave-one-out methods?
- What are the pros and cons for Bootstrapping?
- What is the stratified k-fold cross validation?

MODEL EVALUATION

MODEL EVALUATION METRICS

Performance measure

We can measure the performance via the following metrics:

- Error rate/accuracy
- Precision/recall/F1 score
- ROC/AUC

Error rate/accuracy

- Given a dataset D, the error rate

$$E(f;D) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}(f(\boldsymbol{x}_i) \neq y_i)$$

- Accuracy

$$acc(f;D) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}(f(\boldsymbol{x}_i) = y_i)$$

$$= 1 - E(f;D)$$

Error rate/accuracy

- Given a distribution $\mathcal{D}$ and a probability density function $p(\cdot)$

$$E(f;\mathcal{D}) = \int_{\boldsymbol{x}\sim\mathcal{D}} \mathbb{I}(f(\boldsymbol{x}) \neq y)p(\boldsymbol{x})d\boldsymbol{x}$$

$$acc(f;\mathcal{D}) = \int_{\boldsymbol{x}\sim\mathcal{D}} \mathbb{I}(f(\boldsymbol{x}) \neq y)p(\boldsymbol{x})d\boldsymbol{x}$$

$$= 1 - E(f;D)$$

Confusion matrix

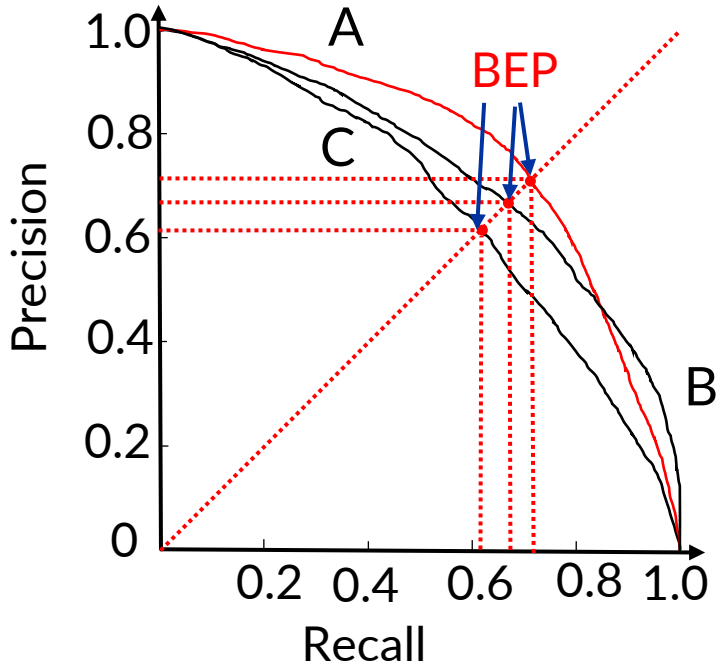
- In binary classification problems, samples can be treated as
  - True positive (TP)
  - False positive (FP)
  - True negative (TN)
  - False negative (FN)
  - TP+FP+TN+FN = # of all samples

- Confusion matrix

		Predicted Class	
		Class=1	Class=0
Actual Class	Class=1	$TP = f_{11}$	$FN = f_{10}$
	Class=0	$FP = f_{01}$	$TN = f_{00}$

Precision and recall

- Precision $P = \frac{TP}{TP + FP}$
- Recall $R = \frac{TP}{TP + FN}$
- P-R curve and Break-Event Point (BEP)



F1-score

- F1-score, or F-score: $$F1 = \frac{2 * P * R}{P + R} = \frac{2 * TP}{\#\ of\ samples + TP - TN}$$

- General form of F-score: $$F_\beta = \frac{(1 + \beta^2) * P * R}{(\beta^2 * P) + R}$$

  - $\beta = 1$: F1-score
  - $\beta > 1$: Recall is more important
  - $0 < \beta < 1$: Precision is more important

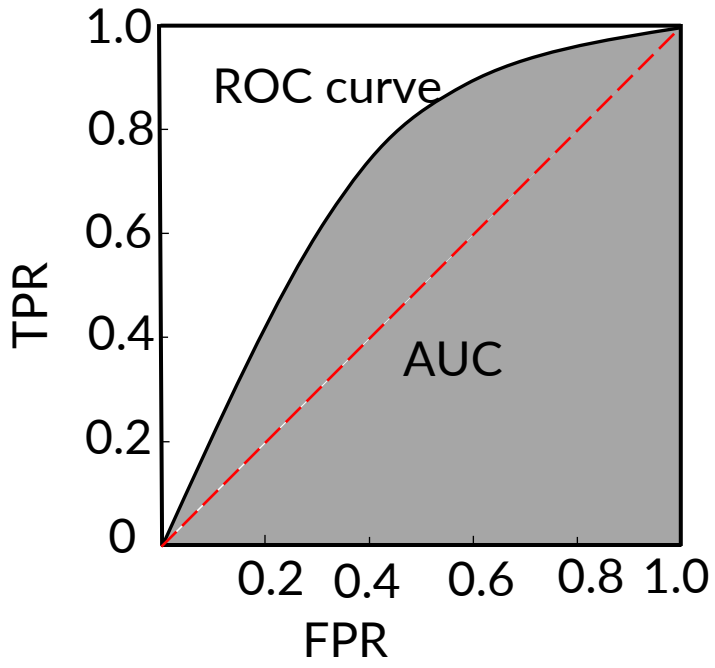
ROC curve

- Receiver Operating Characteristic (ROC) curve
- True positive rate (TPR)

$$TPR = \frac{TP}{TP + FN}$$

- False positive rate (FPR)

$$FPR = \frac{FP}{TN + FP}$$



Area Under Curve (AUC)

- When two ROC curves intersect, AUC is used to evaluate the performances of two classifiers

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

Group Discussion

- In what scenario we should use F1 score rather than accuracy?
- In what scenario we should use precision and recall?
- Why do we use AUC as an evaluation metric?

CONCLUSION

- Introduce the process of model training:
  - Model validation.
  - Hyper-parameter tuning.
  - Training error v.s. Generalization error.
  - Underfitting v.s. Overfitting.

- Introduce the model evaluation methods and metrics:
  - Methods: hold-out, cross-validation, and bootstrapping.
  - Metrics: error rate/accuracy, precision/recall/f1, ROC/AUC.