2.1 How to evaluate and choose a model?

The difference between **the actual prediction result of the sample** by the learner and **the true value of the sample** is called error. We can divide error into three types as belows:

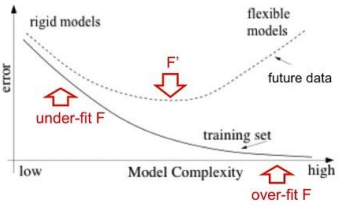
- Errors on the training set are called training errors or empirical errors.
- Errors on the test set are called test errors.
- Errors on all new samples is called the generalization error.

Overfitting and Underfitting

Obviously, all what we want is a good learner on the new sample, a learner with a small generalization error. However, we define the following phenomena according to the ability of a learner:

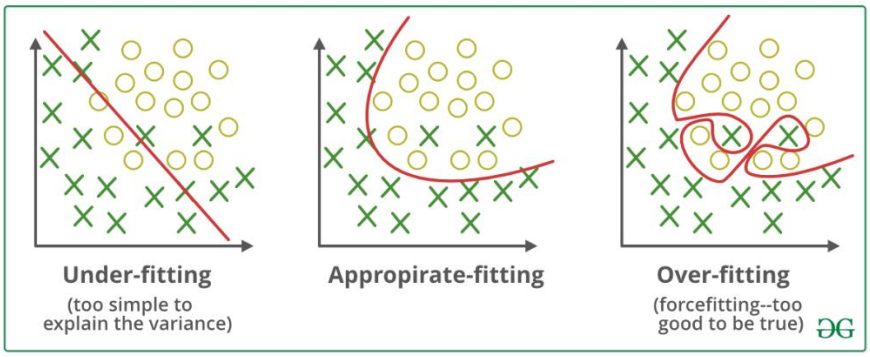
- the learning ability is so strong that the **unusual characteristics contained in the training samples are learned**, which is called **"overfitting"**.
- the learning ability is so poor that the **general properties of the training samples are not well learned**, which is called **"underfitting"**.

- Over-fitting:
  - predictor too complex (flexible)
    - fits "noise" in the training data
    - patterns that will not re-appear
  - predictor *F* over-fits the data if:
    - we can find another predictor *F'*
    - which makes more mistakes on training data: $E_{train}(F') > E_{train}(F)$
    - but fewer mistakes on unseen future data : $E_{gen}(F') < E_{gen}(F)$

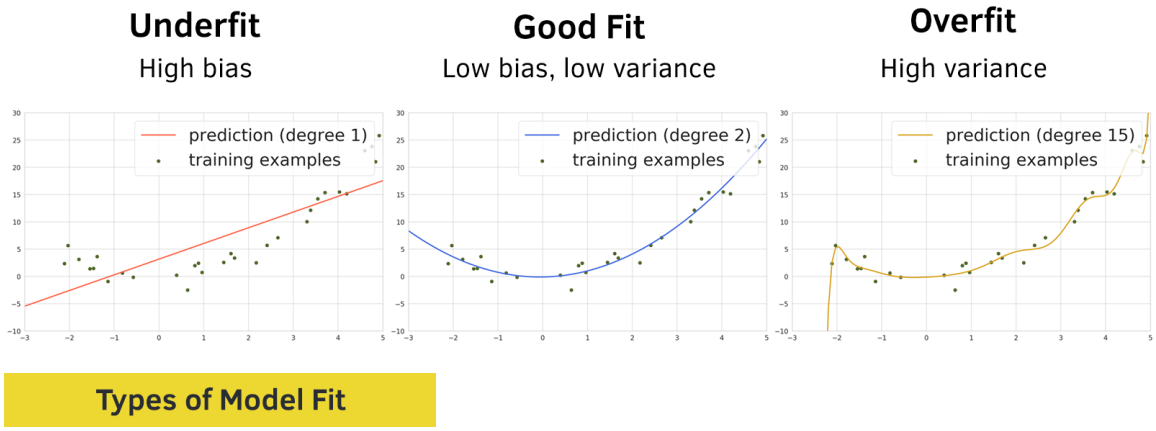


- Under-fitting:
  - predictor too simplistic (too rigid)
  - not powerful enough to capture salient patterns in data
  - can find another predictor *F'* with smaller E_{train} and E_{gen}

Examples for Overfitting and Underfitting
E.g. 1:



E.g. 2:



E.g. 3:

2.2 Evaluation Methods

As mentioned above, we want to get a low generalization error learner, and the ideal solution is to evaluate the generalization error of the model and choose the one with the lowest generalization error. However, the generalization error refers to the applicability of the model to all new samples, so we cannot directly obtain the generalization error. Therefore, we **usually use a "test set" to test the learner's ability to distinguish new samples**, and then take the "test error" on the "test set" as the approximation of the "generalization error". Obviously, **the test set we choose should be as mutually exclusive as possible with the training set**. Here's a short story to explain why:
If the teacher gives 10 exercises for students to practice, and the teacher uses these same 10 questions

as the exam questions, it is obvious that some students can only do these 10 exercises and get high marks. It is obvious that the exam results cannot effectively reflect the real level. Therefore, the teacher should use 10 similar but not totally same questions as exams questions.

Similarly, we want to test a model if it has good generalization performance, just like teacher want to test students if they learn well in the course. For our question, we can get the following table as a comparison.

Test Student	Test Model
Exam	Test Process
Exam Questions	Test Set
Practical Questions	Practice Set

2.2.1 Division for training set and test set

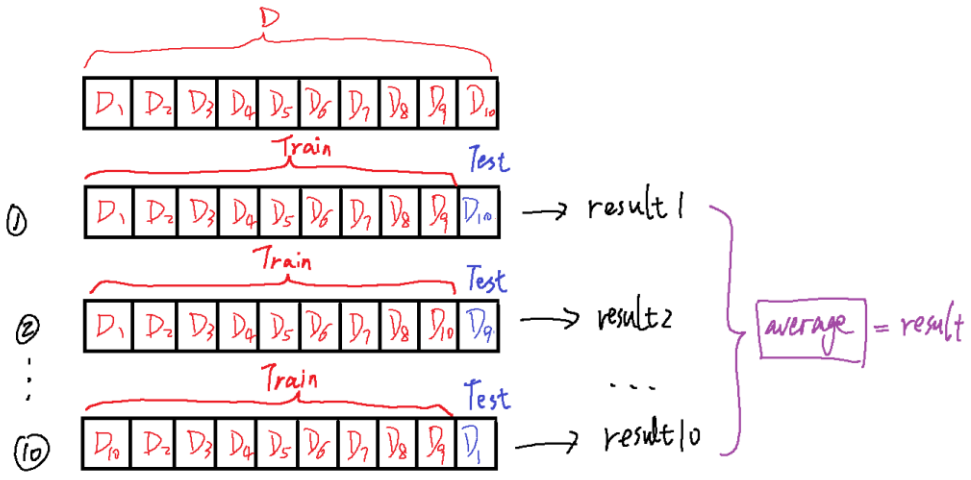
As mentioned above, we want to use “test set” as “new samples”, which means we use the "test error" of a "test set" as an approximation of the "generalization error", so **effectively dividing the initial data set into mutually exclusive "training set" and "test set"** would be significant. The following are some commonly used classification methods:

2.2.1.1 Hold Out Method

Data set D is divided into two mutually exclusive sets, one as training set S and the other as test set T, satisfying $D = S \cup T$ and $S \cap T = \emptyset$. The common classification is that **approximately 2/3-4/5 samples are used for training, and the rest are used for testing**. At the same time, due to the randomness of the division, the results of the single set aside method are often not stable enough, so multiple random divisions and repeated experiments are usually used to take the mean value.

2.2.1.2 Cross Validation Method

The data set D is divided into k mutually exclusive subsets of the same size, satisfying $D = D1 \cup D2 \cup ... \cup Dk$, $Di \cap Dj = \emptyset$ $(i \neq j)$. The idea of the cross validation method is that the union set of **k-1 subsets is used as the training set and the remaining subset is used as the test set**, so that there are k kinds of training set/test set partition, so that k times of training and testing can be carried out, and finally the mean value of the test results of k times can be returned. The usually used k for "k-fold cross validation" is 10, which is shown in the figure below.



Especially when there is only one sample in each subset of k subsets, it is called "leave one method". Obviously, the evaluation result of leave one method is relatively accurate, but it also takes so long for the computer to compute.

2.2.1.3 Bootstrapping Method

Given the data set D containing m samples, **randomly select one sample from D each time, copy it into D', and then put the sample back into the initial data set D and shuffle**, so that each sample can be collected in the same probability. Repeat m times to get the data set D' containing m samples. It's easy to know the limit of the probability that samples will not be collected in m samples is:

$$\lim_{m\to\infty}\left(1-\frac{1}{m}\right)^{m} \rightarrow \frac{1}{e} \approx 0.368$$

2.2.1.4 Comparison of division methods

	Hold Out Method	Cross Validation Method	Bootstrapping Method
Suitable	dataset is big enough	dataset is big enough	dataset is not easy to be divided
Size	train set < whole dataset	train set < whole dataset	train set = whole dataset
Pick	each can be picked as train set	each can be picked as train set	36.8% maybe not be picked as train set

2.2.2 Adjust Parameters

With different parameter configurations, the performance of the learning model is often significantly different. This is commonly referred to as "parameter adjustment" or "parameter tuning". A common practice is to select a range and step size for each parameter to make the learning process feasible. For example, given that the algorithm has 3 parameters and only 5 candidate values are considered for each parameter, there will be 5×5×5=125 models to examine for each training/test set.

2.2.3 Performance Measure

Performance measure is an evaluation standard to measure the generalization ability of models.

2.2.3.1 MSE / Error Rate / Accuracy

Mean Square Error (MSE)

In the regression task, that is, the problem of predicting continuous values, the most commonly used performance measurement is "mean squared error". Many classical algorithms use Mean Square Error (MSE) as the evaluation function:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)$$

Notes:

- n: numbers of samples
- Y_i: actual value
- $\hat{Y}_i$: predict values

Error rate and Accuracy

In the classification task, that is, the prediction of discrete values, error rate and accuracy are most commonly used. **Error rate** refers to the proportion of the sample number of classification errors in the total number of samples. **Accuracy** refers to the proportion of the sample number of classification errors in the total number of samples, easy to know "error rate + accuracy =1".

2.2.3.2 Accuracy Rate / Precision / Recall rate / F-Score

Example:

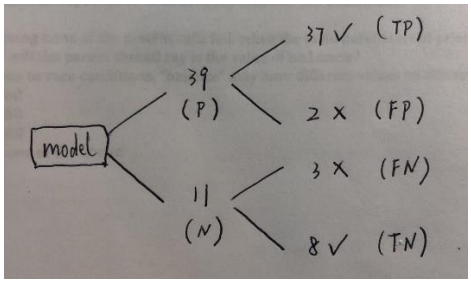
Assume the following problem scenario: There are 50 students in a class, 40 pass and 10 fail in an exam. It is now necessary to predict all passing students based on some characteristics. To understand the meaning of these indicators, we first need to understand two types of samples:

Positive sample: a sample that belongs to a class. In this case, it's a passing student.

Negative sample: a sample that does not fall into this category. In this case, the student who failed.

Model Outcome:

One model was implemented and given 39 people, 37 students indeed passed, and the other two actually failed.



Convert Graph to Table:

Prediction	Actual（Judge True or False）	
	Positive (40)	Negative (10)
Be searched（39 are seemed as PASS）	True Positive（37）	False Positive（2）
Not be searched（11 are seemed as FAIL）	False Negative（3）	True Negative（8）

- TP：Be searched as positive, they are actually also positive (Correct prediction)
  In this case: be searched PASS, actually PASS
- FP：Be searched as negative, they are actually also negative (Type 1 error)
  In this case: be searched PASS, actually FAIL
- FN：Not be searched as positive, they are actually also positive (Type 2 error)
  In this case: be predicted FAIL, actually PASS
- TN：Not be searched as positive, they are actually also positive (Correct prediction)
  In this case: be predicted FAIL, actually PASS

Computation:

	Actual Positive	Actual Negative
Predicted Positive	True Positive(TP)	False Positive(FP) (Type 1 Error)
Predicted Negative	False Negative(FN) (Type 2 Error)	True Negative(TN)

$$\text{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total Population}}$$

$$\text{Error Rate/Misclassification rate} = \frac{\text{False Positive + False Negative}}{\text{Total Population}}$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{Predicted Positive(TP+FP)}}$$

$$\text{Sensitivity/Recall} = \frac{\text{True Positive}}{\text{Actual Positive(TP+FN)}}$$

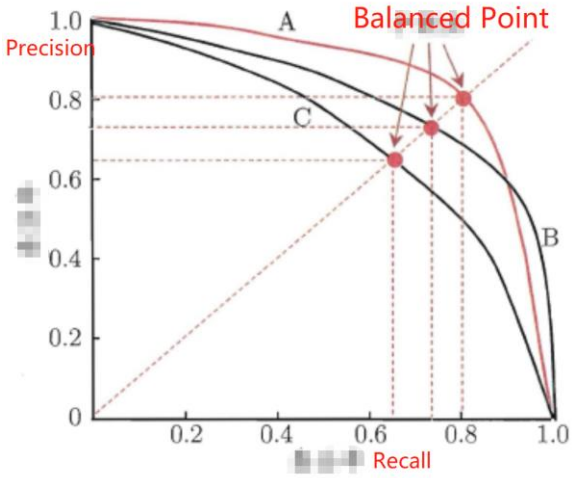
$$\text{Specificity} = \frac{\text{True Negative}}{\text{Actual Negative(FP+TN)}}$$

$$\text{F1 Score} = \frac{2 * (\text{Recall} * \text{Precision})}{\text{Recall + Precision}}$$

[1] Accuracy = no. of sample with correct classification / no. of all samples
In this case, accuracy = (37 + 8) / 50 = 90%
[2] False Rate = no. of sample with wrong classification / no. of all samples
In this case, accuracy = (2 + 3) / 50 = 10%
[3] Precision = no. of sample with correct classification / no. of all predicted positive samples
In this case, accuracy = 37 / 39 = 94.9%
[4] Recall = no. of sample with correct classification / no. of all actual positive samples
In this case, accuracy = 37 / 40 = 92.5%

2.2.3.3 F-Score and P-R Curve

Accuracy and recall are a pair of contradictory measures. For example, if we want to make as much more users as possible interested in the content we push, we can only push one content we think might be the most interesting, thus missing some content that users are interested in and the recall rate is low. If want to push all content that the user is interested in, that has to push on all content only, so accurate rate is low.



How do you evaluate the P-R curve? If the P-R curve of one learner A is completely enveloped by the P-R curve of another learner B, then the performance of B is better than that of A. If the curves of A and B intersect, **who has the greater area under the curve has the better performance.** However, in general, it is difficult to estimate the area under the curve, when P=R, Balanced Point which has the higher value has the better the performance.

Sometimes we hope to refer to precision and recall at the same time, so a new index f-score is introduced to comprehensively consider Precision and Recall. Where β is used to adjust the weights. When $\beta = 1$, both weights are the same, abbreviated as F1-score. If you think Precision is more important, reduce the value of β, If Recall is more important, then add more.

$$FS = \frac{(1+\beta^2)\cdot(Precision + Recall)}{\beta^2\cdot(Precision + Recall)}$$

2.2.3.4 ROC and AUC

What is ROC?

Receiver operating characteristic curve (ROC) is a comprehensive indicator reflecting sensitivity and specificity.

How to Draw ROC Curve?

If we have obtained the probability output of all samples (the probability belonging to positive samples), then we **rank each test sample by the probability that it belongs to a positive sample**. The following figure is an example with a total of 20 test samples. The column "Class" represents the true label of each test sample (p represents positive sample, n represents negative sample), and "Score" represents the probability that each test sample belongs to positive sample.

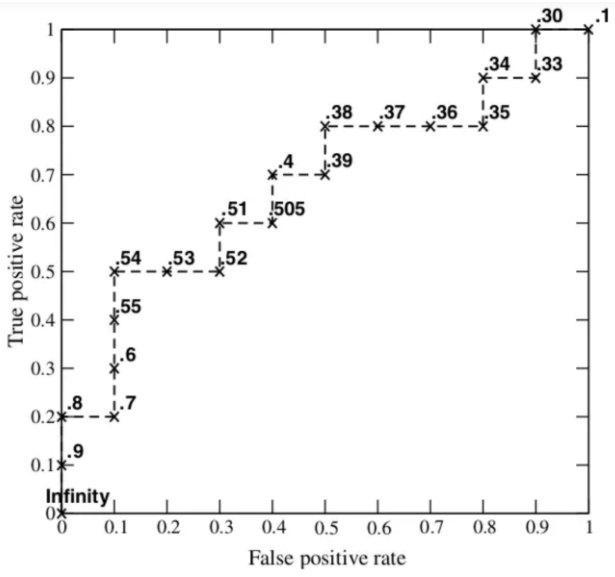
Inst#	Class	Score	Inst#	Class	Score
1	**p**	.9	11	**p**	.4
2	**p**	.8	12	**n**	.39
3	**n**	.7	13	**p**	.38
4	**p**	.6	14	**n**	.37
5	**p**	.55	15	**n**	.36
6	**p**	.54	16	**n**	.35
7	**n**	.53	17	**p**	.34
8	**n**	.52	18	**n**	.33
9	**p**	.51	19	**p**	.30
10	**n**	.505	20	**n**	.1

Next, we take the "Score" value as the threshold from high to low. When the probability of the test sample belonging to a positive sample is greater than or equal to this threshold, we consider it as a positive sample; otherwise, it is a negative sample. For example, for the fourth sample, its "Score" value is 0.6, then samples 1, 2, 3 and 4 are considered positive samples, while other samples are considered negative samples. Therefore, we get:

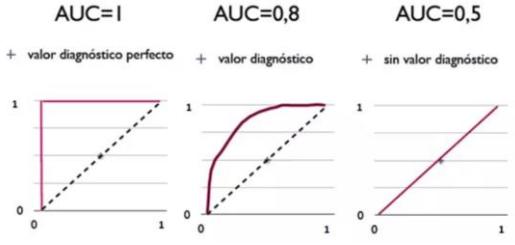
TPR = true positive / acutual positive = 3 / 10 = 0.3

FPR = false positive / acutual negative = 1 / 10 = 0.1

Each time we select a different threshold, we can get a set of (FPR, TPR), a point on the ROC curve. In this way, we obtained a total of 20 sets of FPR and TPR values, and drew them on the ROC curve as follows:

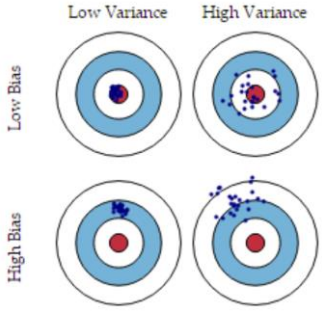


What is AUC?
AUC is area under the ROC curve. The higher AUC is, the better performance of classifier is.



2.2.3.5 Bias and Variance

Bias and variance are two aspects used to measure the generalization error of a model.
Bias: difference between the expected value predicted by the model and the real value, which is used to describe the fitting ability of the model. For example, the real model is a quadratic function, while we assume the model is a first function, which will lead to the increase of deviation (underfitting).
Variance: sum of the difference squares between the expected value of the model and predicted value, which is used to describe the stability of the model. For example, the real model is a simple quadratic function, while we assume the model is a higher-order function, which will lead to the increase of variance (overfitting).



Generalization Error
In supervised learning, the generalization error of the model can be decomposed into the sum of deviation, variance and noise.

$$Err(x) = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$

Given a learning task, ①when the training is insufficient, the fitting ability of the model is not enough and the deviation dominates the generalization error of the model. ②With the progress of training, the fitting ability of the model is enhanced and the variance gradually dominates the generalization error of the model. ③When the training is sufficient, the fitting ability of the model is too strong, then the overfitting occurs (the non-global features of the training data are also learned by the model).