李宏毅深度学习 p8

where does the error come from

A more complex model dose not always lead to better performance on testing data.

error due to "bias" and due to "variance"

Bias and Variance of Estimator

· Estimate the mean of a variable x

　· assume the mean of x is μ

　· assume the variance of x is σ^2

· Estimate of mean μ

　· sample N points $\{x^1, x^2, ..., x^N\}$

$$m = \frac{1}{N}\sum_n X^N \neq \mu$$

$$E[m] = E[\frac{1}{N}\sum_N x^N] = \frac{1}{N}\sum_N E[x^N] = \mu$$

$$Var[m] = \frac{\sigma^2}{N}$$

　　variance depends on the number of samples

· Estimator of variance σ^2

　· samples N points $\{x^1, x^2, \cdots, x^N\}$

$$m = \frac{1}{N}\sum_N x^N$$

$$s^2 = \frac{1}{N}\sum_N (x^N - m)^2$$

· Biased estimator

$$E[s^2] = \frac{N-1}{N}\sigma^2 \neq \sigma^2$$

解释

以打靶为例，Bias就是你瞄偏了，Variance是子弹偏移了你瞄准的地方。

simpler model is less influenced by the sample data.

　Bias

$$E[f^*]=\bar{f}$$

　· Bias If we average all the f^*, it is close to $\breve{a}$

式子越复杂，Bias越小，Variance越大。

Variance过大称为overfitting

Bias过大称为underfitting

What to do with large Bias

· **Diagnosis**

　　· If your model cannot even fit the training examples,then you have large Bias(underfitting)

　　· If you can fit the training data,but large error on testing data,then ou brobably have large Variance(overfitting)

· **For Bias**

　　· Add more features as input

　　· A more complex model

· **For Variance**

　　· More data (very effective,but not always practical)　　　　可以generate假的data

　　· Regulariztion(因为只包含那些较平滑的曲线所以可能伤害Bias)

Model Selection

· There is usually a trade-off between bias and variance

· Select a model that balances two kinds of error to minimize total error

· What you shoul NOT do:根据用Training Set训练出来的model在Testing Set上的准确率来选择model,因为用自己的Testing Data选择的model不一定在真实的Testing Data上效果也很好，通常而言效果是比较差的。
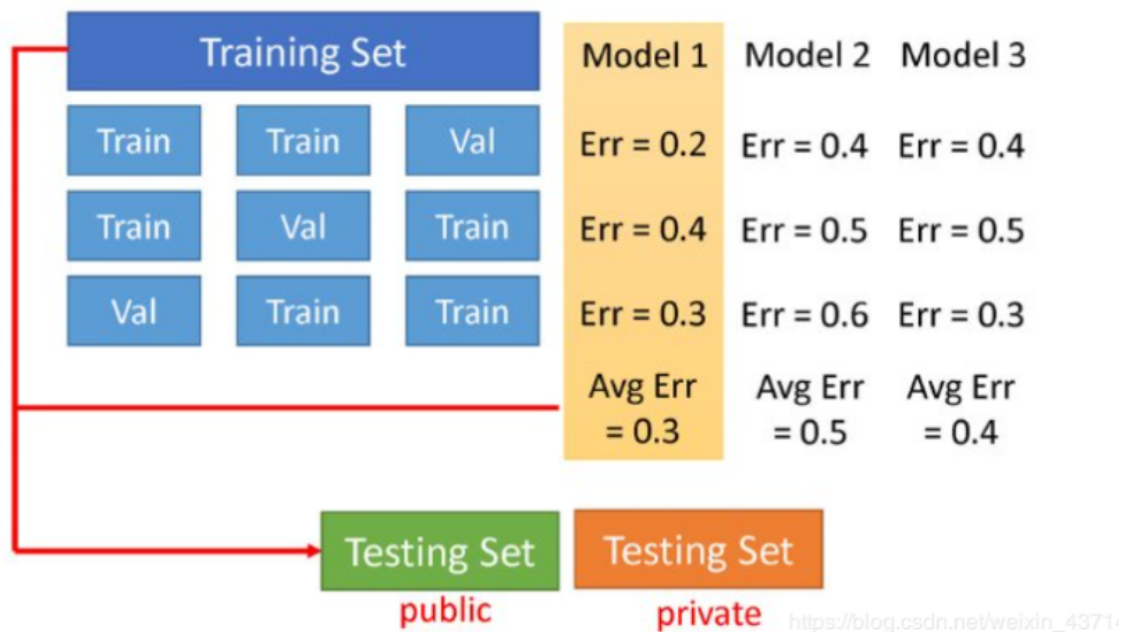
那么我们应该怎么呢?

Cross Validation 交叉验证

　　方法就是把原先的Training Set分为Training Set和Validation Set两部分，根据model在validation上的误差来选择最优model。注意：不建议在得出Training Set的误差后，再回去调整model(这样做相当于把Testing Set也当作Training Set的一部分了)

　　也可以用N折交叉验证的方式选择模型，如下图所示

N-fold Cross Validation

	Model 1	Model 2	Model 3
	Err = 0.2	Err = 0.4	Err = 0.4
	Err = 0.4	Err = 0.5	Err = 0.5
	Err = 0.3	Err = 0.6	Err = 0.3
	Avg Err = 0.3	Avg Err = 0.5	Avg Err = 0.4

Training Set: Train | Train | Val / Train | Val | Train / Val | Train | Train

Testing Set **public** — Testing Set **private**

注意：原则上，少用public Testing Set来调整model，这样往往在private Testing Set上得到的结果与在public Testing Set上得到的记过差距是比较小的。