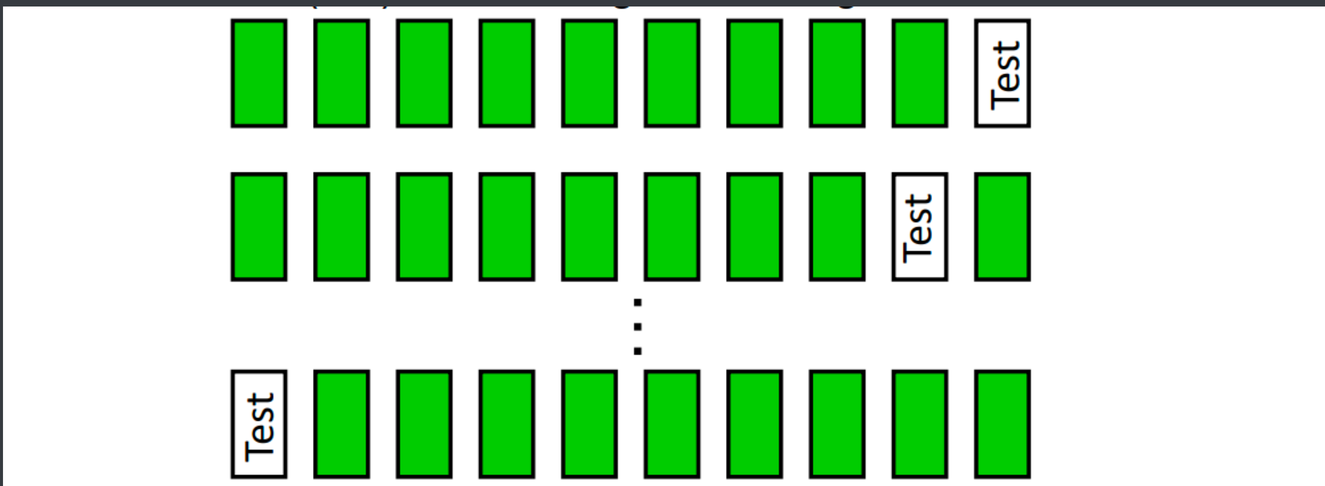
experimental methodology

generalizing

- test set, training set
- cross validation: make the most of data

  example:

Training Folds	Testing Fold	Correct	Incorrect	Accuracy
S1, S2, S3, S4	S5	35	15	0.70
S1, S2, S3, S5	S4	30	20	0.60
S1, S2, S4, S5	S3	37	13	0.74
S1, S3, S4, S5	S2	33	17	0.66
S2, S3, S4, S5	S1	40	10	0.80

  but one calss might be completely missing in the training data

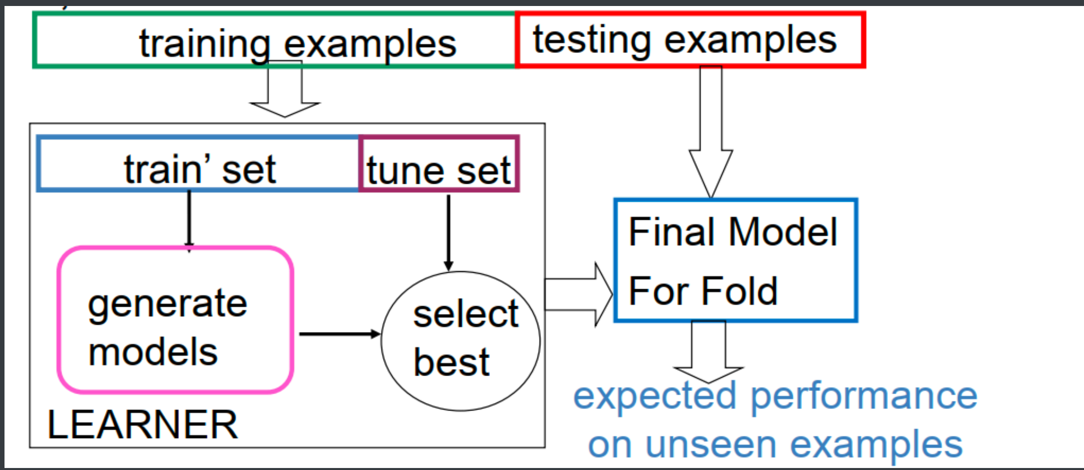
  problem: an infrequent class might be missing at all

  -> solution: stratification, then concatenate all test results
- leave-one-out cross validation

  leave only one but not a set out

  very computationally expensive
- big pitfall: test data is not necessarily representative of real data!

Tuning

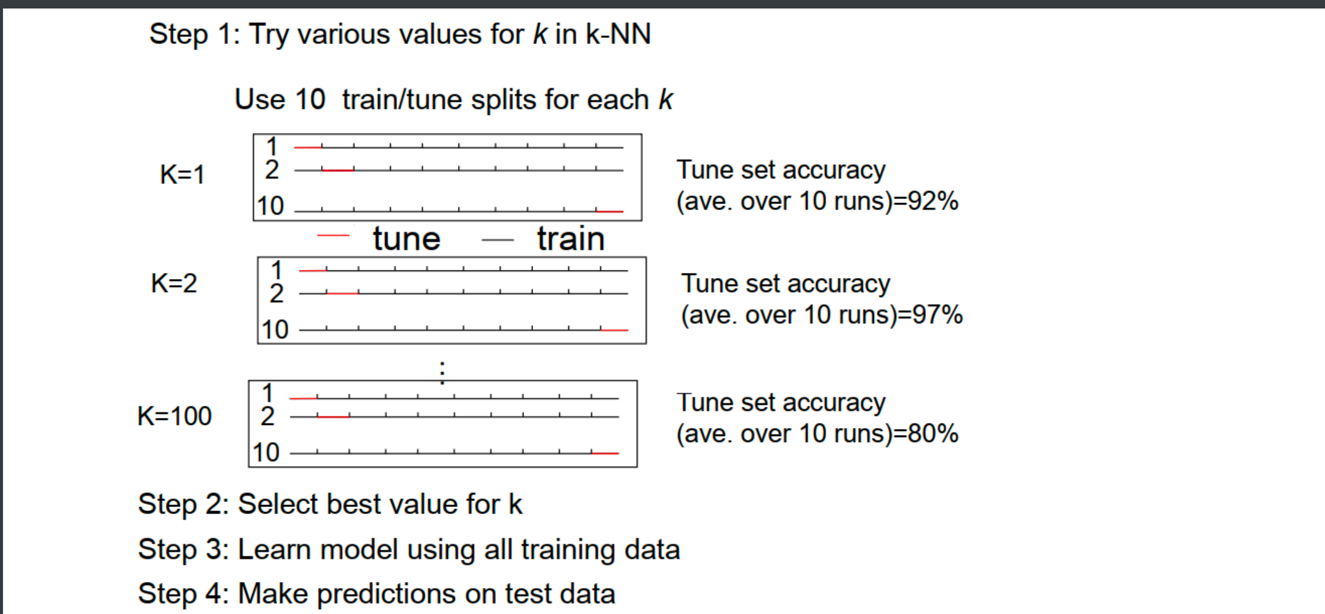


- internal cross-validation

  motivation:single tuning set can be unreliable

  - CV on training data to select the parameters
  - Score setting by average performance
  - Pick best one
  - Train model on all training data

  example:



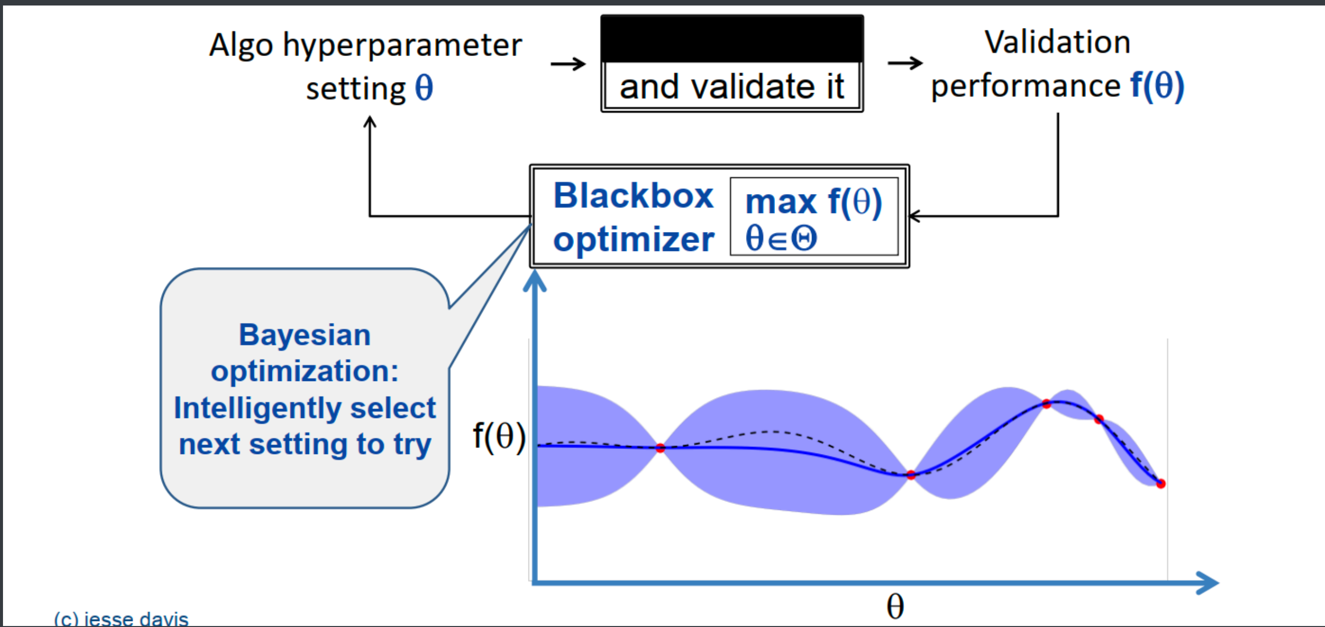
- spread of values: wide-range, 避免等差数列 （tune之前的参数设置需要拉开差距）
- hyperparameter search

  search problem

random search (sample possible values) <-> grid search

random search 通常很有效，因为存在一些效果差不多的configurations，grid search效率很低

automated hyperparameter tuning

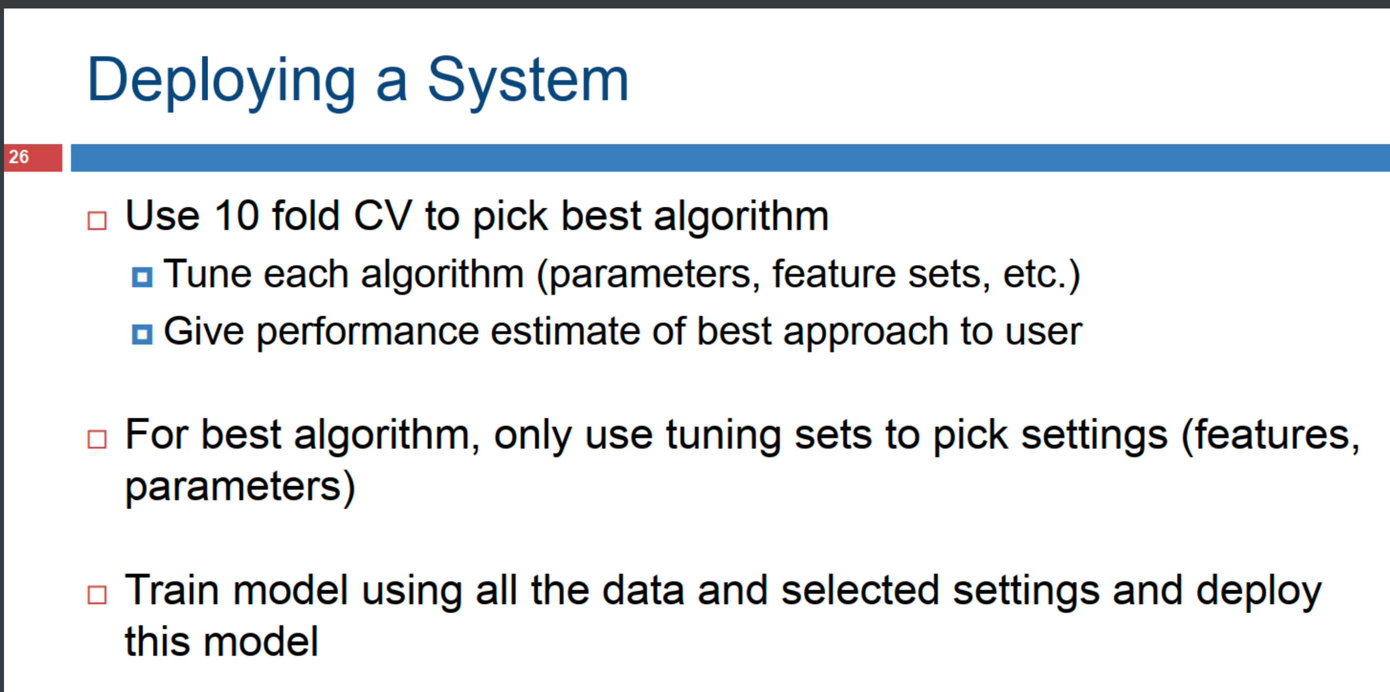


红色的点是采样点，第二个紫色包中可能存在更好的点

- parameter-less algirithm:

  internally sets all parameters, rather than user explicitly setting the hyperparameters

to wrap up,



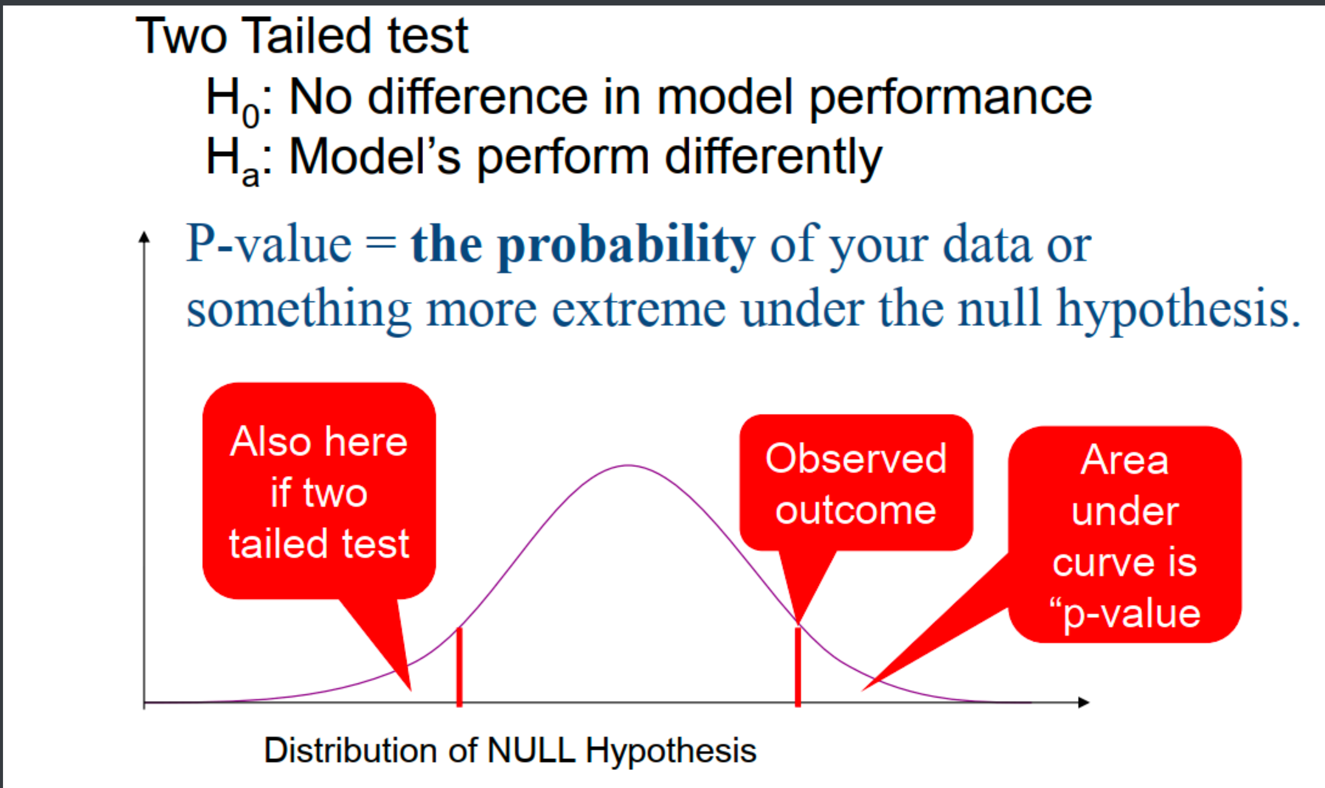
comparisons

- 假设检验 hypothesis testing：

  null hypothesis 原假设 H_0 - 两个模型performance没有显著差异

  alternative hypothesis 备选假设 H_a - 两个模型performance有显著差异

  计算统计量和p值，比较p值和显著性水平；



- McNemar's Test

  core: how often was $f_1$right and f_2 wrong on the sample example, or the other way around?

	f_1 right	f_1 wrong
f_2 right	*A*	*B*
f_2 wrong	*C*	*D*

  to see if B and C are significantly different:

$$\chi^2 = \frac{(b-c)^2}{(b+c)},$$

- paired t-test

	Accuracies on Testsets				
Algo L1:	88	69	79	…	72
Algo L2:	79	62	74	…	75
δ :	+9	+7	+5	…	-3

null hypothesis: algos have equivalent average accuracies

if p-value is low, then reject null hypothesis.

assumptions of the t-test:

1. test statistic is normally distributed

   e.g. if metric is classifier accuracy, then ✅

   AUC ROC then ❌ because usually not normally distributed

   -> wilcoxon signed-rank test: 非参数检验方法，不需要正态性假设

2. independent sample of test-examples

   cross validation通常违反这一条规则，但实际上t检验对独立性假设的偏差是非常robust的；

- sign test

  记录两个算法L1和L2在测试样本上“意见不一致”情况下的“win”次数，M是较大的win次数；

  假设两个模型性能相似，用二项分布b(N, 0.5)计算L1或者L2在随机情况下赢至少M次的概率，如果概率很低则拒绝假设；

- wilcoxon signed rank test - on multiple data sets

RF	DT	RF-DT	|RF-DT|	Rank	Signed Rank
90	88	2	2	2	2
99	66	33	33	10	10
65	75	-10	10	6	-6
67	55	12	12	7	7
87	73	14	14	8	8
69	75	-6	6	4	-4
73	72	1	1	1	1
74	89	-15	15	9	-9
80	85	-5	5	3	-3
77	70	7	7	5	5

$W_+ = 33$ and $W_- = 22$

$T_{Wilcox} = \min(33, 22) = 22$

calculation:

step 1 compute performance difference

step 2 rank the absolute differences

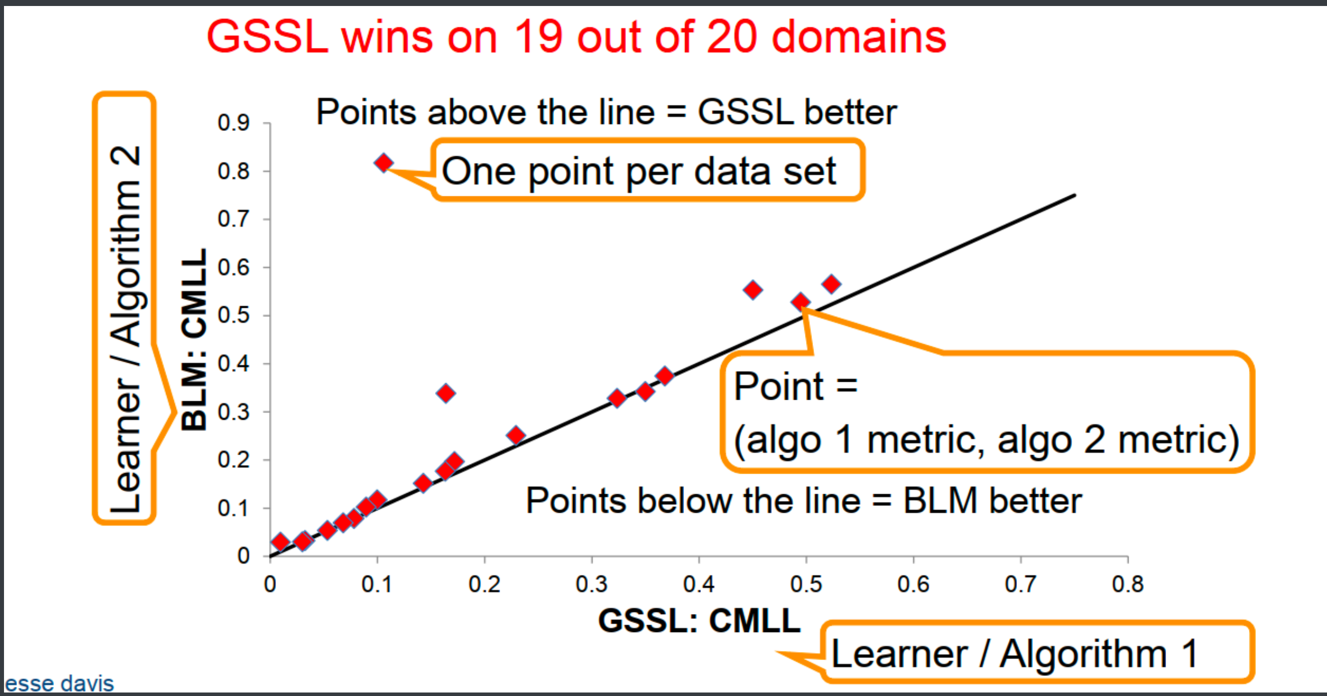
step 3 add sign to rank

step 4 sum of positive ranks & sum of negative ranks, take the minimum value

meaning:

generallyW值越大说明两个模型差异越大，但需要和P值配合来跟显著性水平比较！

- graphical comparison



一个数据点代表一个数据集上的结果，黑色直线代表两个模型性能相同