Review

Error Metrics

- why misclassification-rate/accuracy is not enough?
- confusion matrix
  - Sensitivity=Recall: TP / P = TP / (TP + FN)
  - Specificity: TN / N = TN (TN + FP)
  - Precision: TP / (TP + FP) (**)
- Bayes' rule

Resampling & Cross-validation

- bootstrap (resample with replacement)
  - sample size: = N (original sample size) (why? The accuracy of statistical estimates depends on the sample size...)
  - the probability that each instance is sampled at least once ≈ 0.632
- cross validation
  - K-fold
  - When the training set size is smaller (smaller k):
    - more variety of fit (the model obtained vary)
    - Error-estimate: high bias low variance

Model selection

Large-/medium-sized dataset

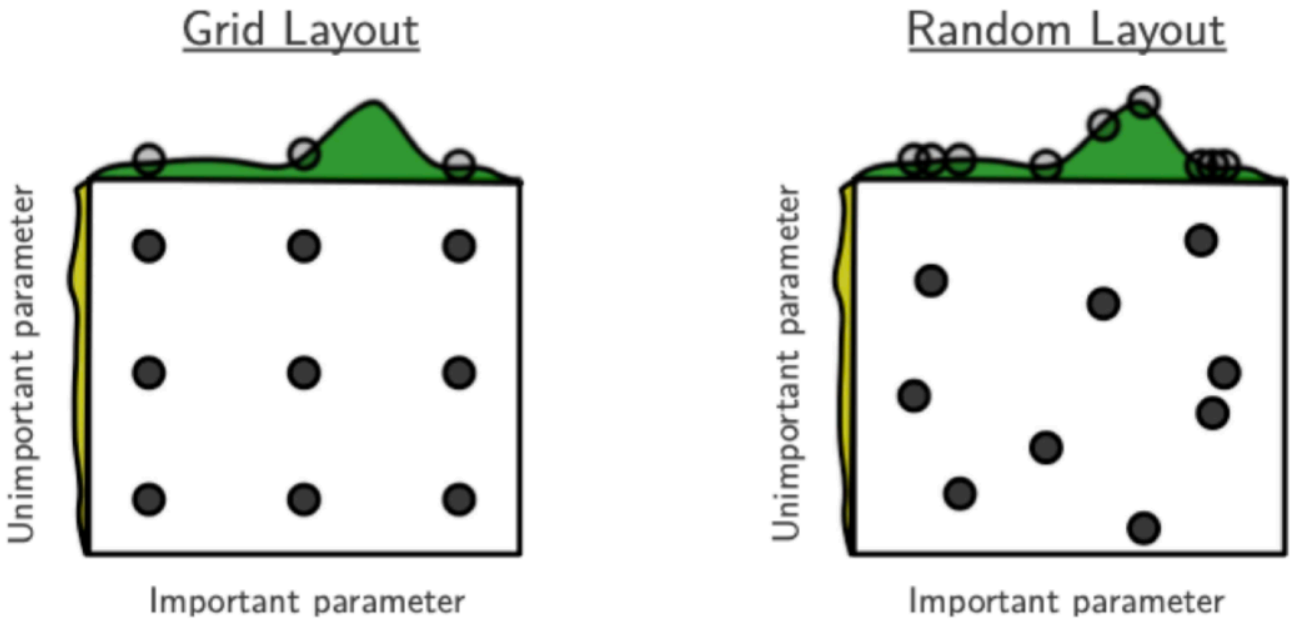
- for each model, test its performance on train&validation set.
  - internet sized datasets: 1 train set, 1 validation set. (method 0)
  - Medium-sized datasets: k-fold cross-validation. (method 1)
- pick the model of the best performance.
- train the best model on all train and validation sets and test it on test set.

Small dataset

- for each model, estimate validation&test error using nested k-fold cross validation
- pick the best model: according to performance on [the mean across trials] of the inner-cross validation folds
- estimate the test error of the best model: outer cross-validation fold performance

Grid Search

- compare randomized & grid search



Statistical Testing

- T-test
- parsimony principle
- when a computationally expensive test is needed...
- highly-specialized v.s. general-purposed algorithms

Discussion Questions

See: https://github.com/COGS118A/DiscussionSection/blob/main/W5_discussion.pdf

A4

KNN:

Euclidean distance: `np.linalg.norm`

Algorithm (to determine the label of x_i)

```
1. compute its distance to all points from the training-set
2. rank the points in the training set from closest to furthest, pick the first-k
3. majority vote: scipy.stats.mode
```

Q: what if there are more than k points are 'k-nearest'?

Q: what if there are more than one 'most frequent' labels?