Resampling Methods

resampling - train a model. shuffle & take another sample to train
repeat many times
model assessment - evaluate model's performance
model selection - selecting proper level of flexibility

Cross validation

1) Validation set approach
split the data equally for training & testing (training set & validation set / hold out set)
the test MSE will be different each time
drawbacks
- validation estimate of test error rate can be highly variable, depending on which observations are in training vs testing set
- since only a subset of observations are included in the training set, the model won't have enough data to learn. hence the validation test MSE might be overestimated, compared to the whole data

2) Leave One Out Cross Validation (LOOCV)
a single observation is used to validate vs $n-1$ training observations. repeating this n times gives n MSE
LOOCV estimate for the test MSE,

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^{n} MSE_i$$

advantages
- less bias. does not overestimate the test error rate
- not much variance in the training / validation set splits

this is an expensive implementation
with linear / polynomial MSE regressions, the shortcut is

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{1 - h_i} \right)^2$$

h_i is leverage. i^{th} residual is divided by the residual
leverage lies b/w $\frac{1}{n}$ & 1. reflects the amount that an observation influences.

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n} (x_{i'} - \bar{x})^2}$$

3) K-Fold cross validation

divide the dataset into k groups. one of them is test set.
repeat the model training for k times, leaving another as test ~~set~~ ~~test~~.
(LOOCV is a variation of k-fold where $K = n$)

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^{k} MSE_i$$

the MSE curves might underestimate / overestimate than the true test test error curve
but we are mainly interested in the point of minimum MSE in the curve. because we might be performing cross validation on a no. of diff statistical models, or a single model with different levels of flexibility.

4) Bias variance tradeoff for k-fold cross validation
test error rate will be more for validation set approach > k fold > LOOCV
from the perspective of bias, LOOCV is preferred to k-fold CV
in LOOCV, each model differs by just 1 observation
error estimates are highly correlated. averaging them → high variance estimate
k fold CV - errors are less correlated. avg them → low variance
there's bias variance tradeoff in choosing the k in k-fold CV.
$k = 5$ or 10 usually gives the good result with a balance of bias, variance

cont^n of Resampling methods

5) Cross validation on classification problems

when it comes to classification, we use no. of mis classified observations instead of MSE.

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} Err_i \qquad Err_i = I(y_i \neq \hat{y}_i)$$

though cross validation error curve slightly underestimates the test error rate, it takes on a minimum very close to the best value of degree of polynomial / k - in case of KNN classifier

Bootstrap

how much does the sample is similar to true population?
whats the variability of the sample statistic (sample mean etc)?
to answer these, we do bootstrapping
where new samples are created from existing dataset with repetitions, the repetitions account for the duplicates, missing value conditions of the true population and explains the variance of sample statistics

α that minimizes the variance

$$\hat{\alpha} = \frac{\hat{\sigma}_y^2 - \hat{\sigma}_{xy}}{\hat{\sigma}_x^2 + \hat{\sigma}_y^2 - 2\hat{\sigma}_{xy}}$$

$\hat{\sigma}_x^2, \hat{\sigma}_y^2 = Var(X), Var(Y)$

$\hat{\sigma}_{xy} = Covar(X, Y)$