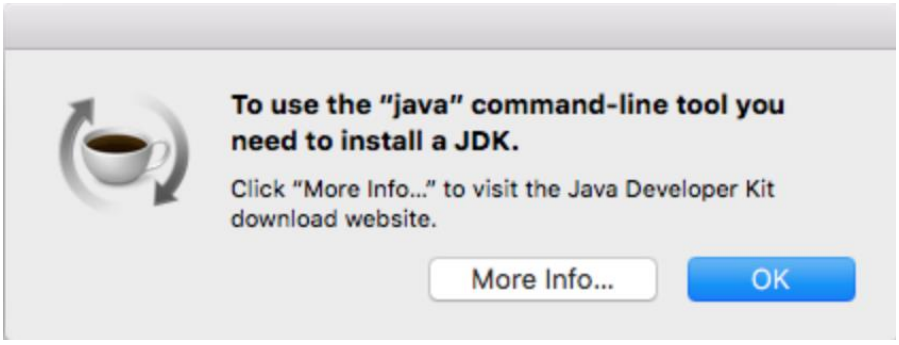
MA679 Lab1

Notes on h2o installation in R

1. H2o official documentation on installation in R: http://docs.h2o.ai/h2o/latest-stable/h2o-docs/downloading.html#install-in-r

2. Install JDK link:
   version 11 https://www.oracle.com/technetwork/java/javase/downloads/jdk11-downloads-5066655.html
   version 13 https://www.oracle.com/technetwork/java/javase/downloads/jdk13-downloads-5672538.html



Notes: If install h2o by using
install.packages("h2o"),
it doesn't install the latest version of h2o, you need to install JDK 7-12.

If install the latest h2o by
install.packages("h2o", type="source", repos=(c("http://h2o-release.s3.amazonaws.com/h2o/latest_stable_R"))),
JDK 13.0.2 (2020-01-14) is fine.

Evaluation Model Metrics

Regression

Notations: n denotes sample size, k denotes number of covariates, $\hat{L}$ denotes maximum likelihood.

- **R2 (R squared) & Adjusted R2**
  - Usage: Goodness of fit.
  - Definition:

$$R^2 = 1 - \frac{RSS}{TSS},$$

$$Adj - R^2 = 1 - \frac{\frac{RSS}{n-k-1}}{\frac{TSS}{n-1}}$$

$$RSS = \sum (y_i - \hat{y}_i)^2,\ TSS = \sum (y_i - \bar{y})^2,$$

- Interpretation: An R_2 of 1 indicates that the regression predictions perfectly fit the data.
- Remarks: Adj-R2 is better since it puts penalty on dimension.

- **MSE (Mean Squared Error)**
  - Usage: measures the average of the squares of the errors or deviations.
  - Definition:

$$MSE = \frac{1}{n}\sum (y_i - \hat{y}_i)^2,$$

  - Remarks: Sensitivity to outliers: RMSE (root mean squared error)>MSE>MAE (mean absolute error)

- **AIC & BIC**
  - Usage: Model selection.
  - Definition:

$$AIC = 2k - 2\ln(\hat{L}),$$
$$BIC = \ln(n)k - 2\ln(\hat{L}),$$

  - Interpretation:
    - Forward feature selection: choose the feature that after adding it, the AIC/BIC is smallest.
    - Backward feature elimination: eliminate the feature that after eliminating it, the AIC/BIC is largest.
  - Remarks: BIC usually gives a smaller set of covariates, since bigger penalty is put on dimension of features when the sample size is large.

Classification

- **Confusion matrix**

		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error
	Predicted condition negative	**False negative**, Type II error	**True negative**

- **Precision, recall, F1 score**

  precision=TP/(TP+FP),
  recall=TP/(TP+FN),
  F1=2*precision*recall/(precision + recall)

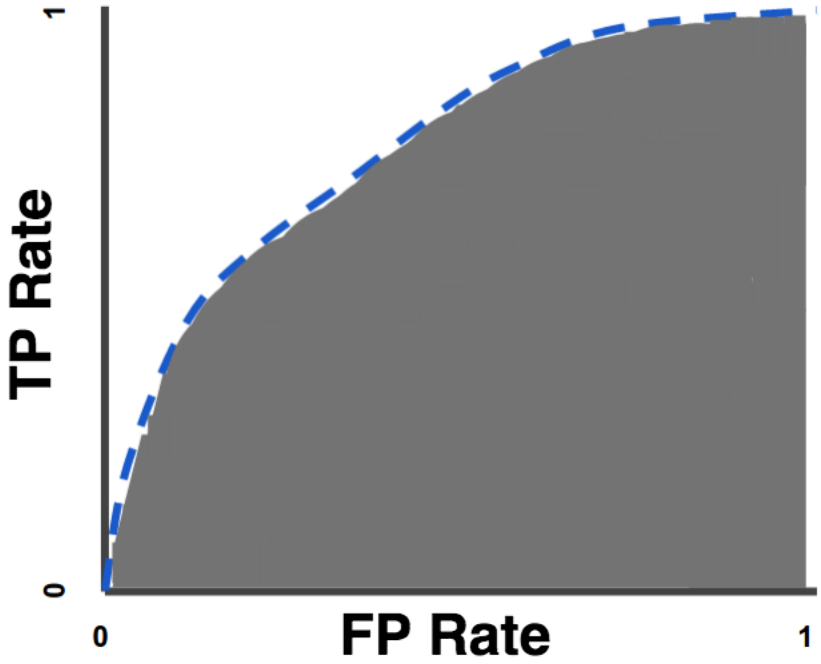
  - Remarks: With an imbalanced data, use these metrics instead of accuracy.

- **Logloss**

  For logistic regression:

$$\boldsymbol{logloss} = -\frac{1}{n}\sum(y_i \ln(p_i) + (1 - y_i)\ln(1 - p_i))$$

- **AUC (Area under ROC curve)**
  - TP rate=TP/(TP+FN), FP rate=FP/(FP+TN)



  - Interpretation: An AUC of 1 indicates a perfect classifier, while an AUC of 0.5 indicates a poor classifier.