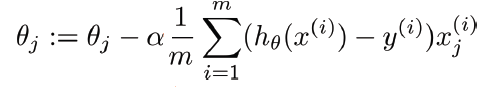
**Regression model 梳理**

Graphical user interface, text, application

Description automatically generated

1. Gradient Descent Based Algorithms
2. OLS regression



1. Distance-Based Algorithms
2. Regression tree
3. Bagging
4. Random forest
5. boosting
6. Tree-Based Algorithms
7. Support vector regressor (SVR)
8. KNN regression
9. 其他（Subset selection等）
10. 小结
11. Gradient Descent Based Algorithms
12. OLS regression

* 原理：Estimate the parameters using least squares, i.e., minimizing the RSS.
* 优势：High interpretability. The parameters in the linear regression model are very easy to interpret. Good for inference purpose.
* 劣势： linearity的假设很可能会与现实有出入。
* 假设：LINE assumption (linearity, independent and normally distributed error terms, homoscedasticity).另外，regression还要求multivariate normality和low multicollinearity.

<https://towardsdatascience.com/all-the-annoying-assumptions-31b55df246c3>

* 预处理要求：
  + 对qualitative predictors要create dummy variables (可以直接用fast dummies package)
  + Multivariate normality的假设要求我们最好纠一下偏度
  + Low multicollinearity的假设要求我们不同时保留高相关性的predictors
* 可能的用途：
  + 可以作为ensemble中的一种prediction方法。
  + 可以用来帮助我们做inference. 通过t test和f test来解释什么是影响房价的最重要因素，也可以对interaction terms做t test, 来解释predictor之间可能会有怎样的synergy effect。
  + 可以为feature selection做参考。先用t-test看哪些predictor是不重要的，然后再对这些predictor做general linear F-test，看能不能一起舍弃掉。

1. Tree Based Algorithms
2. regression tree

* 原理：recursive binary splitting & cost complexity pruning
* 优势：比OLS更适用于non-linear data; high interpretability; 不用创dummy variables
* 劣势：too short-sighted, prediction accuracy可能偏低

1. Bagging

* 原理：bootstrap -> recursive binary splitting -> average
* 可以用variable importance measure plot来interpret
* 劣势：different trees may be collineared

1. Random forest

* 原理：bagging + randomly select m predictors at each split
* 优势：
  + High prediction accuracy
  + Can handle many predictor variables
  + 不需要做缺失值处理：Maintains accuracy even when a large proportion of the data is missing
  + 可以用variable importance measure来interpret每个predictor的重要性
* 劣势
  + Can overfit datasets that are particularly noisy
  + Interpretability有一定局限性：For data including categorical predictor variables with different number of levels, random forests are biased in favor of those predictors with more levels. Therefore, the variable importance scores from random forest are not always reliable for this type of data
* 假设：No formal distributional assumptions, random forests are non-parametric and can thus handle skewed and multi-modal data as well as categorical data that are ordinal or non-ordinal.

https://bccvl.org.au/algorithms-exposed-random-forest/#:~:text=ASSUMPTIONS,  
are%20ordinal%20or%20non%2Dordinal

* 预处理要求
  + 理论上不需要填补缺失值，不需要纠偏度，不需要创dummy variable, 不需要做feature scaling
  + 但是如果结合PCA的话，需要做feature scaling
  + 最好处理一下collinearity, 因为collinearity虽然不会影响prediction accuracy，但是会影响interpretability

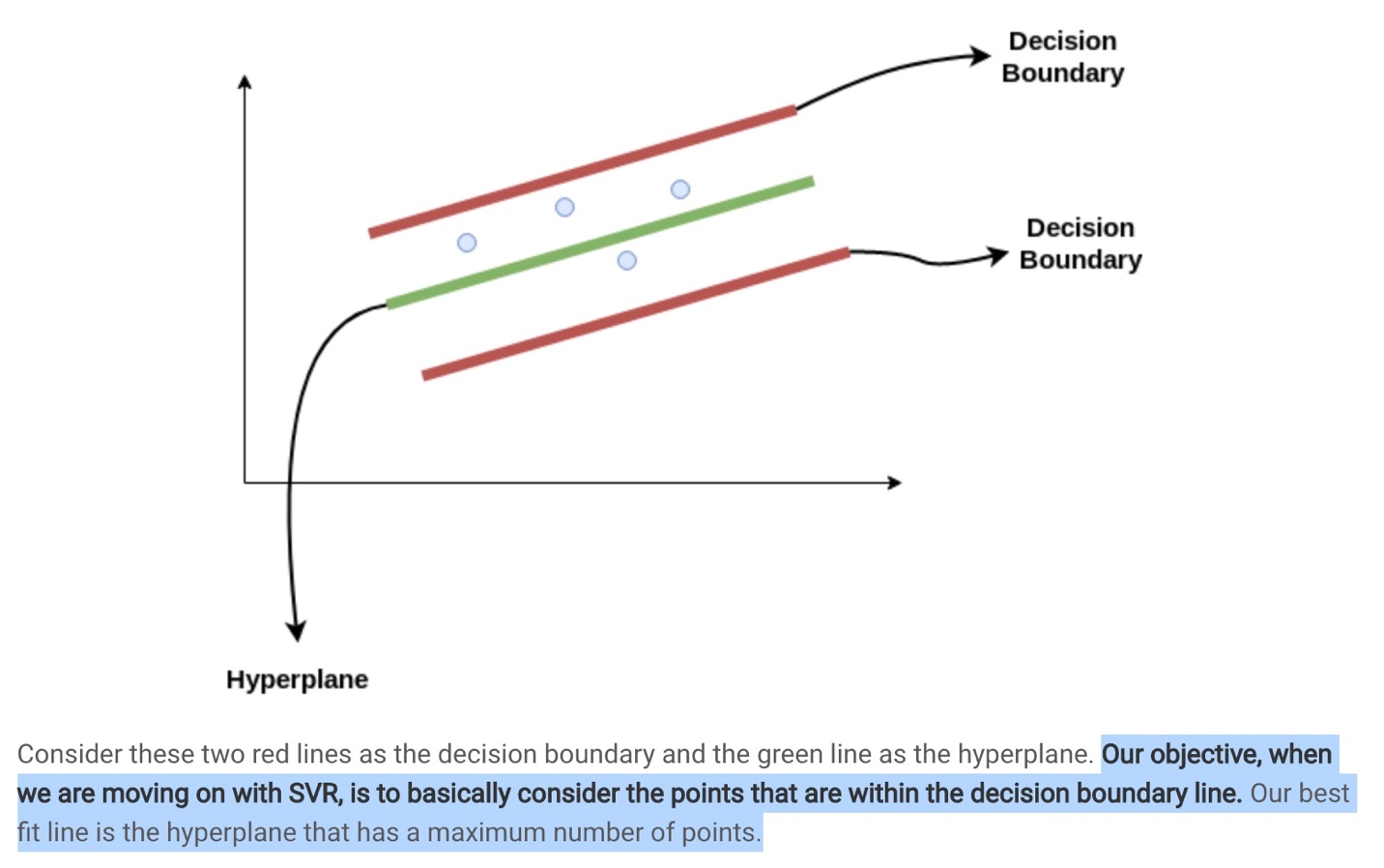
<https://www.researchgate.net/post/Are-Random-Forests-affected-by-multi-collinearity-between-features>

* 用R语言做的random forest实操：<https://www.kaggle.com/patrycjawalicka/r-house-prices-decision-tree-random-forest-pca>

1. Boosting

* 原理：growing trees sequentially by using information from previously grown trees; learns slowly, allowing different shaped trees to attack the residuals; use CV to select the number of trees B, also need to be careful about lambda and d.

1. Distance Based Algorithms
2. Support Vector Regressor (SVR)

* 原理：
* 优势：acknowledges the presence of non-linearity in the data
* 假设：nonparametric method, 对数据没有特殊的要求，也没有独立的假设
* 预处理要求：kaggle上的notebook进行了缺失值处理和standardization
* 实操：<https://www.kaggle.com/himaoka/house-simple-svr-support-vector-regression>

1. KNN regression

* 原理：Non-parametric method, K controls the flexibility

1. 其他：

* Subset selection methods (best subset/forward stepwise/backward stepwise), regularization methods (ridge and lasso), dimension reduction method (PCA and PLS)主要是用来balance bias and variance的，没有各自的假设。唯一要注意的是PCA隐含假设了variance大的predictor的predictive power要大一些，而且PCA需要做feature scaling.
* bootstrap可以用来估计最终预测结果的方差，也可以用来计算population parameter的bootstrap percentile confidence interval.
* Feature Scaling
  + Gradient Descent Based Algorithms和Distance-Based Algorithms需要feature scaling, Tree-Based Algorithms不需要
  + Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors, Neural Networks, Random Forest, and SVR.
  + Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

<https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/>

1. 小结：
2. 预处理

* handle missing value
  + 降低bias的普遍要求
  + Random forest不需要填补缺失值
* feature scaling (standardization/normalization)
  + SVR和OLS regression需要
  + PCA需要feature scaling
  + Random forest不需要
* feature selection (zero variance)
  + 降低noise的普遍要求
* fix skewed features
  + OLS需要 (normality assumption)
  + Random forest不需要纠偏
* fix collinearity
  + collinearity会影响OLS和Random Forest’s variable importance measure的interpretability
* (create dummy variables)
  + OLS需要
  + Random forest不需要
* outliers?
  + Normalization既可以实现feature scaling，也可以纠正outlier.
* Add new features?

1. 模型选择