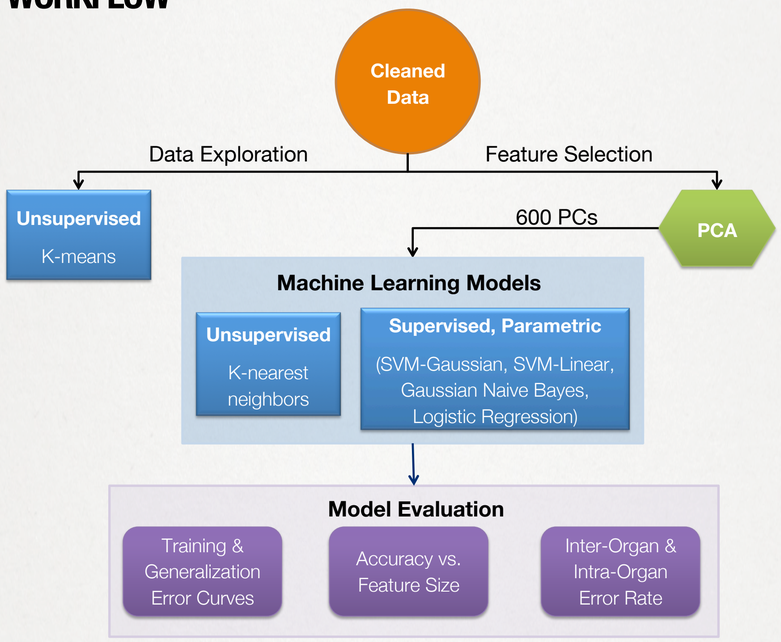
MACHINE LEARNING FOR CLASSIFICATION OF LUNG AND KIDNEY CANCER TYPE

Vivek Jain | Weizhuang Zhou | Yifei Men

INTRODUCTION

Changes in DNA methylation profile are known to cause variable gene expression observed in cancer cells.
We demonstrate that whole-genome methylation profile can be used to build effective models to discriminate between lung and kidney cancer and their sub-types using machine learning techniques.

WORKFLOW



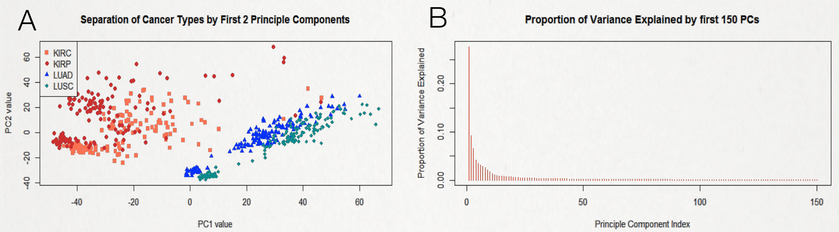
DATA

150 methylation profiles for each of the following 4 cancer types from The Cancer Genome Atlas:

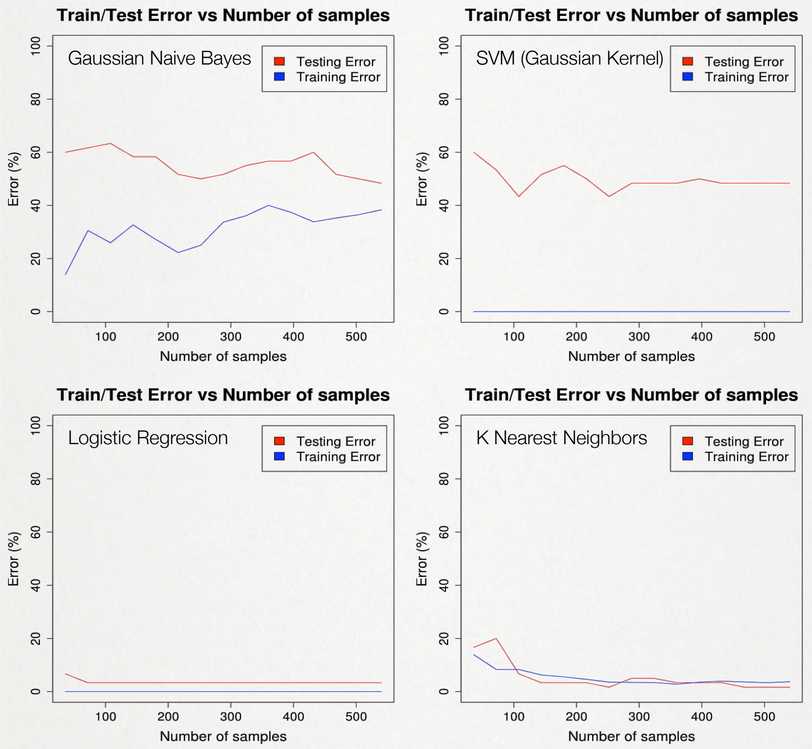
- Kidney renal clear cell carcinoma (KIRC)
- Kidney renal papillary cell carcinoma (KIRP)
- Lung adenocarcinoma (LUAD)
- Lung squamous cell carcinoma (LUSC)

Methylation was assessed at 485,577 positions across the genome and reported as numerical values. Single value imputation and quantile normalization were performed to correct missing values and batch effects in experimentation.

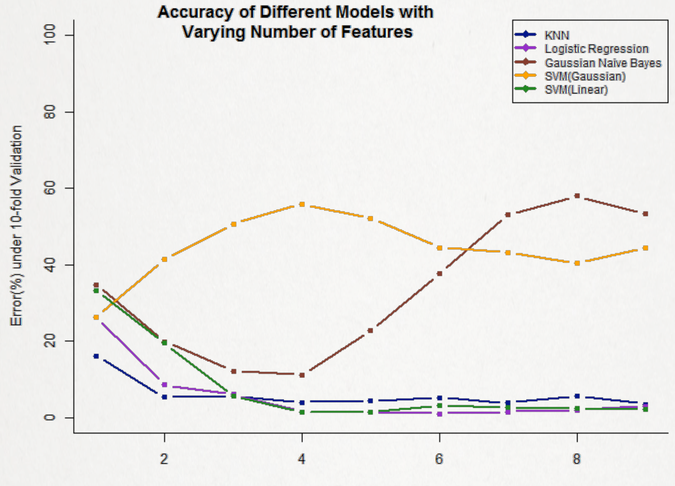
RESULTS



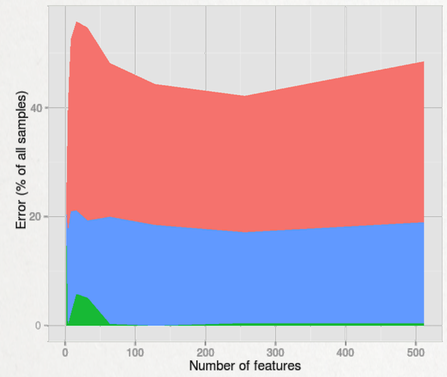
Feature selection using PCA (A) Cancers from different organs are well-separated spatially using the first 2 principle components. (B) The first few principle components explain a large proportion of variance observed in data.



Training and Generalization (Testing) Error of Different Models. Gaussian Naive Bayes model demonstrate high training and generalization error, suggesting high bias. Gaussian SVM has high generalization but low training error, suggesting over-fitting. Logistic regression and Linear SVM (not shown) have similar profiles, indicating that performance can be further improved with larger sample size. Performance of KNN model is promising, but does not improve with increasing number of samples.



Accuracy of models with varying PCA feature size. Logistic Regression, Linear SVM and KNN performed well with increasing number of features. Non-linear parametric models (Gaussian Naive Bayes and Gaussian SVM) demonstrate increasing error rates with higher number of features, likely due to increased over-fitting.



Dissection of Errors in Gaussian Naive Bayes Model. Majority of errors are Between Organ and Within Lung misclassifications. The breakdown of errors did not change significantly with change in number of features.

CONCLUSION

- Logistic regression and linear SVM accurately classify (>95%) the given 4 types of cancers with low bias and variance.
- Effective classification can be achieved by restricting feature space to less than 100 top principle components.
- Future work includes identifying contribution of individual methylation sites using coefficients of logistic regression, and uncovering their biological relevance.