Drug Sensitivity Estimation Based on Adaptive Empirical Conditional Expectation

FAN Min†

†Department of Mathematics, Hong Kong University of Science and Technology

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Abstract

Predicting drug sensitivity to specific cancer cell line is an essential part for personalized cancer therapy. Traditional machine learning algorithms do not perform very well in this scenario. The reasons probably is because the features of cancer cell lines are binary and the number of training samples is limited due to the high cost of experiments. In this work, we proposed a new model called Adaptive Empirical Conditional Expectation(AECE) from the perspective of probability and statistics. The model achieved the highest performance among traditional machine learning algorithms, like Support Vector Regression(SVR), Random Forest(RF), Gradient Boosting(GB), Multilayer Perceptron(MLP). Moreover, the model is tuning-free, capable to candle missing features and does not need to assume training and test cell lines are from the same distribution.

Problem Setting

In training data, there are 542 cancer cell lines. Each cell line has 60 binary features, $X = [x_1, \ldots, x_{60}] \in \{0,1\}^{60}$, which correspond to the mutation status of 60 cancer genes. The drug sensitivity of cancer cell lines to the specific drug, Afatinib, is measured by IC50s which is expected to be smaller for high sensitive cell lines and larger for low sensitive cell lines. Roughly speaking, the negative values of logarithmic IC50 indicate the cell line is sensitive to the drug and positive values indicate the resistance to the drug. In testing data, we need to predict IC50s of 100 cancer cell lines with the mutation status of 60 cancer genes.

Empirical Conditional Expectation

Given feature X, IC50, denoted as $y(X)$, can be regarded as a random variable (r.v.) due to the randomness in biology. In testing data, the estimator (prediction) of IC50 is also a r.v., denoted as $\hat{y}(X)$, since estimator is a function of IC50s in training data, which are also r.v.s.

In practice, we minimize mean squared error (MSE):

$$\begin{aligned}\text{MSE} &:= \frac{1}{n}\sum_{i=1}^{n}(\hat{y}(X_i) - y(X_i))^2 \\ &= \sum_{j=1}^{t}\frac{n_j}{n}\left\{\frac{1}{n_j}\sum_{i=1}^{n_j}[\hat{y}(X_j) - y(X_j)]^2\right\}\end{aligned}$$

In theory, by law of large number, minimizing MSE is equivalent to minimize following:

$$\text{Overall Error} := \mathbb{E}_{X\sim p_{test}}[\mathbb{E}_y(\hat{y}(X) - y(X))^2].$$

One essential assumption of most machine learning algorithms is $p_{test} = p_{train}$. However, it may not hold sometimes, like personalized cancer therapy of specific patient. We can see it later that our model can also tackle this case.

Theorem Suppose $\mathbb{E}\,X^2 < \infty$ and $\mathcal{F}$ is a σ-algebra,

$$\mathbb{E}\,(X|\mathcal{F}) = \arg\min_{Y\in\mathcal{F}} \mathbb{E}\,(X-Y)^2.$$

It tells us that $\mathbb{E}\,(X|\mathcal{F})$ is the best guess of X given the information $\mathcal{F}$. We can implement the same idea to minimize the red part above, that is, $\mathbb{E}_Y(\hat{y}(X) - y(X))^2$.

First, we construct $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \ldots \subset \mathcal{F}_{60}$ by computing p-values of features. In particular, for feature i, we can test the hypothesis using T-Test

$$H_0 : \mathbb{E}(y|X_i = 0) = \mathbb{E}(y|X_i = 1) \quad \text{versus} \quad H_1 : \mathbb{E}(y|X_i = 0) \neq \mathbb{E}(y|X_i = 1).$$

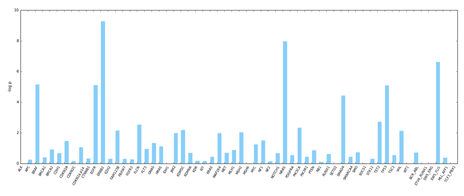

Fig. 1: P-values

And we sort features as p-value descending order. $\mathcal{F}_k$ is the feature space discarding last $60-k$ insignificant features. Then given specific feature $\bar{X}$we can approximate $\mathbb{E}\,(y(X = \bar{X})|\mathcal{F}_k)$ by averaging $y(X)$ of training samples with the same feature of $\bar{X}$ in $\mathcal{F}_k$, called empirical conditional expectation.

For example, in the figure below, $\mathcal{F}_2$ is the feature space containing feature 1 and 2. We need to predict IC50s of the sample in last row. $\mathbb{E}\,(y(X = \bar{X})|\mathcal{F}_2) = \frac{1}{2}(2.1 + 2.2) = 2.15$.

IC50s	Feature 1	Feature 2	Feature 3	Feature 4
2.1	1	0	0	0
3.1	1	1	1	0
2.2	1	0	0	1
-1	0	0	1	1
?	1	0	1	0

Fig. 2: Example of Empirical Conditional Expectation

The larger k is, the more information $\mathcal{F}_k$ contains and the more accurate $\mathbb{E}\,(y(X = \bar{X})|\mathcal{F}_k)$ is. So we can propose Vanilla Empirical Conditional Expectation(ECE) below.

Algorithm 1 Empirical Conditional Expectation

Input:
Training samples with label $\{X^i, y^i(X)\}_{i=1}^n$, test sample $\bar{X}$.
Output:
Predicted value for test $\hat{y}(\bar{X})$

1: Run T-test for each feature in train data, sort those features by p-values and construct $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \ldots \subset \mathcal{F}_{60}$ using p-values.
2: Set $k = 60$.
3: Find the samples $\{X_k^i, y_k^i(X)\}_{i=1}^{n_k}$ in train data which have the same features with $\bar{X}$ in $\mathcal{F}_k$.
4: If $n_k = 0$, $k := k-1$ and go to step 3, otherwise, go to step 5.
5: **return** return $\hat{y}(\bar{X}) = \frac{1}{n_k}\sum_{i=1}^{n_k} y_k^i(X)$.

Adaptive Empirical Conditional Expectation

$$\begin{aligned}\mathbb{E}_y(\hat{y}(X) - y(X))^2 &= \mathbb{E}_y(\hat{y}(X) - \mathbb{E}\,\hat{y}(X))^2 + \mathbb{E}_y(\mathbb{E}\,\hat{y}(X) - \mathbb{E}\,y(X))^2 + \mathbb{E}\,(\mathbb{E}\,y(X) - y(X))^2 \\ &= \text{Var}(\hat{y}(X)) + \text{Bias}^2(\hat{y}(X)) + \text{Var}(y(X)).\end{aligned}$$

Last term $\text{Var}(y(X))$ caused by randomness in biology is intrinsic and we can only minimize the first two terms. If $\hat{y}(\bar{X}) = \frac{1}{n_k}\sum_{i=1}^{n_k} y_k^i(X)$, intuitively, the larger k is, the smaller $\text{Var}(\hat{y}(X)$ is and the larger $\text{Bias}^2(\hat{y}(X))$. So we have to make a trade-off on k to minimize the sum ot them.

$\text{Bias}^2(\hat{y}(X))$: $\mathbb{E}\,y(X)$ is approximated by the average of returns in algorithms 1 by bootstrapping 50 times, denoted as $\hat{\mathbb{E}}y(X)$. Given k, $\mathbb{E}\,\hat{y}(X)$ is approximated by the average of empirical conditional expectation with $\mathcal{F}_k$ by bootstrapping 50 times, denoted as $\hat{\mathbb{E}}\hat{y}_k(X)$. Hence, $\text{Bias}^2(\hat{y}(X))$ can be approximated by $(\hat{\mathbb{E}}\hat{y}_k(X) - \hat{\mathbb{E}}y(X))^2$, denoted as $\hat{\text{Bias}}^2(\hat{y}_k(X))$.

$\text{Var}(\hat{y}(X))$: Given k, $\text{Var}(\hat{y}(X)$ is approximated by the variance of empirical conditional expectation with $\mathcal{F}_k$ by bootstrapping 50 times, denoted as $\hat{\text{Var}}(\hat{y}_k(X))$.

k is decreasing from 60. Adopting early stopping idea, we choose the largest k such that $\hat{\text{Bias}}^2(\hat{y}_k(X)) + \hat{\text{Var}}(\hat{y}_k(X)) < \hat{\text{Bias}}^2(\hat{y}_{k-1}(X)) + \hat{\text{Var}}(\hat{y}_{k-1}(X))$.

Adaptive Empirical Conditional Expectation(AECE) is stated as below.

Algorithm 2 Adaptive Empirical Conditional Expectation

Input:
Training samples with label $\{X^i, y^i(X)\}_{i=1}^n$, test sample $\bar{X}$.
Output:
Predicted value for test $\hat{y}(\bar{X})$

1: Run T-test for each feature in train data, sort those features by p-values and construct $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \ldots \subset \mathcal{F}_{60}$ using p-values.
2: Bootstrap from training samples with equal sample size 50 times and each time, implement Algorithm 1 and the return is $\hat{y}_{\text{ECE}}(\bar{X})$. Take the average of $\hat{y}_{\text{ECE}}(\bar{X})$ as the approximation of $\mathbb{E}y(X)$, denoted as $\hat{\mathbb{E}}y(X)$.
3: Set $k = 60$.
4: Bootstrap from training samples with equal sample size 50 times, denoted as $\{X_m^i, y_m^i(X)\}_{i=1}^n$. In time m,
   1. Find the samples $\left\{X_{k,m}^i, y_{k,m}^i(X)\right\}_{i=1}^{n_k}$ in new train data with the same features with $\bar{X}$ in $\mathcal{F}_k$.
   2. if $n_k = 0$, re-bootstrap and go to step 4.1, otherwise, go to step 4.3
   3. $\hat{y}^m(\bar{X}) = \frac{1}{n_k}\sum_{i=1}^{n_k} y_{k,m}^i(X)$.
5: Compute $\hat{\text{Bias}}^2(\hat{y}_k(X))$ and $\hat{\text{Var}}(\hat{y}_k(X))$ using $\hat{\mathbb{E}}y(X)$ and $\{\hat{y}^m(\bar{X})\}_{m=1}^{50}$.
6: If $\hat{\text{Bias}}^2(\hat{y}_k(X)) + \hat{\text{Var}}(\hat{y}_k(X)) < \hat{\text{Bias}}^2(\hat{y}_{k-1}(X)) + \hat{\text{Var}}(\hat{y}_{k-1}(X))$, go to step 7, otherwise, $k := k-1$ and go to step 4.
7: **return** return $\hat{y}(\bar{X}) = \frac{1}{n_k}\sum_{i=1}^{n_k} y_k^i(X)$.

Result

AECE achieved the highest performance in the competition and the MSE is only 3.04446.

#	Team Name	Kernel	Team Members	Score
1	proteusfan			3.04808
	Your Best Entry ↑			
	Your submission scored 3.04446, which is an improvement of your previous score of 3.04808. Great job!			
2	Artie			3.06307
3	tcloaa			3.06693
4	skydog			3.06968
5	thedog			3.07587

Discussion

Advantage:

- capable to candle missing features
- tuning-free
- capable to do prediction of sample with specific feature($p_{train} \neq p_{test}$)

Disadvantage:

- time-consuming in some degree

Potential Improvement:

- use local FDR to replace t-test
- approximate $\text{Var}(\hat{y}(X))$ and $\text{Bias}^2(\hat{y}(X))$ more accurate