Conceptual Guidelines for Model Selection Based on VC Theory

Joon Hyun Byon

June 30, 2016

1. Introduction

In the search for appropriate performance levels, junior data science practitioners, may find themselves in a never-ending spiral down a rabbit hole, especially when a performance ceiling is reached after deploying their favorite benchmark algorithm. Frustration and unchecked creativity, may spark the implementation of esoteric combinations of algorithms in the hope of generating a model that can obtain significant marginal improvements. Although these efforts to search and select the best model can pay off, it has been observed that the process is approached more as an art, than science. For example, junior practitioners often default to a Random Forest, as the first algorithm applied to a new data set, without accounting for the sample size at hand, implement every possibly known algorithm, or may go on rampage to derive as many new variables as time allows, under the assumption that their models will correctly identify the most relevant ones. Although the latter is true for models which have built-in feature selection, the crux of the matter is that many important decisions seem to be made without an appropriate degree of theoretical justification.

According to the literature, the challenge of finding a good performing model can be decomposed into a) finding a small training error and b) assuring that performance obtained during training generalizes well across previously unseen data. Because it can be trivial to satisfy the former by overfitting, the main concern in machine learning is the latter, which this paper attempts to address by providing some conceptual guidelines for model selection using VC Theory. This paper heavily relies on the textbook, Learning from Data (Abu-Mostafa et al., 2012.), where textbook plots have been replicated and problem sets answered to convey key ideas. The paper is divided into two main parts. The first part synthesizes the theory of generalization by presenting the utility of Hoeffding Inequality and VC Inequality and the second part presents some conceptual applications that can help guide decision making during the model selection process.

2. Generalization Theory

2.1 Applying the Hoeffding Inequality to evaluate learning with a single hypothesis or a finite set of hypotheses

Hoeffding Inequality provides an upper bound on the probability that the generalization error, the difference between the sample and population frequency of error, will be greater than a specified level of tolerance. In the case of Bernoulli random variables, it can be states that

$P[|v - u| > \epsilon] \leq 2e^{-2\epsilon^2 N}$, for any sample size N, and any $\epsilon > 0$,

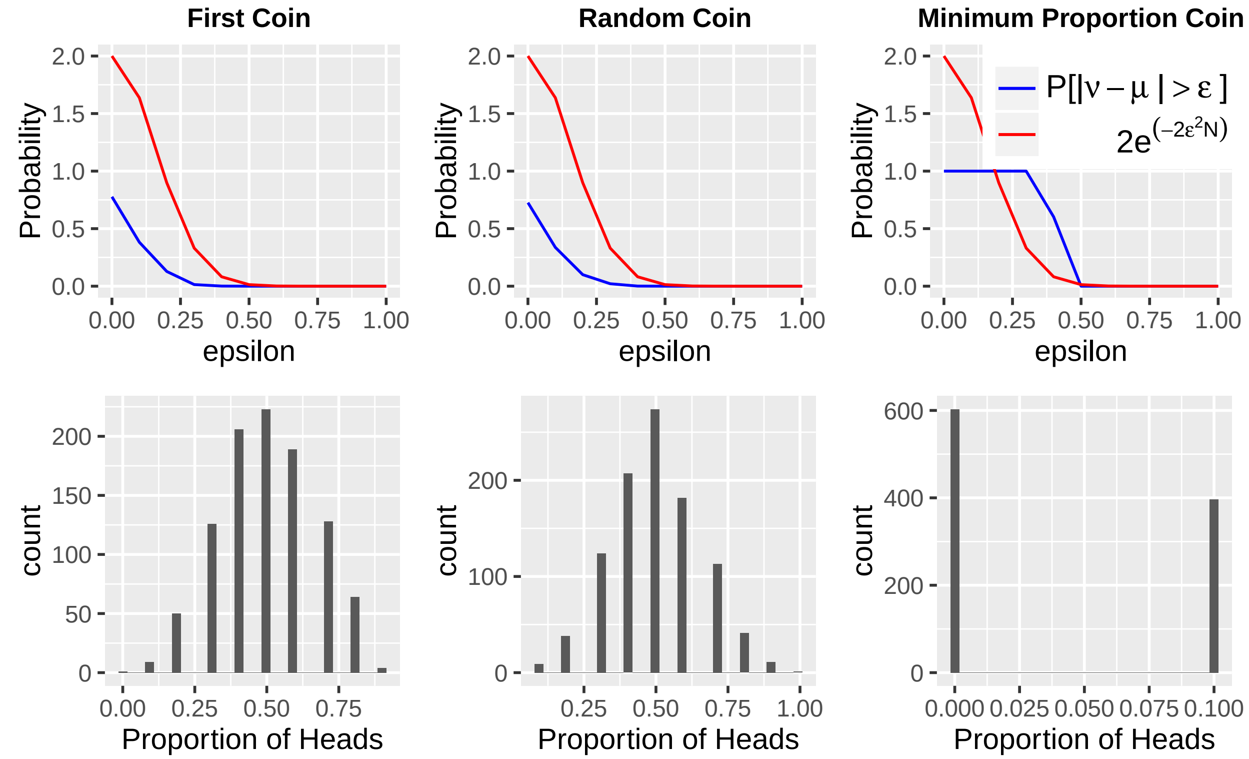
where v is the observed sample error, u is the unknown population error, and ϵ is our tolerance level. Therefore, the Hoeffding Inequality provides us with a probabilistic framework to infer u from v, and as N increases, the more confident we can be on the likelihood of their convergence, where the key assumption is that v is random (Abu-Mostafa et al., 2012, p. 21-22).

In the context of machine learning, we can substitute v and u with the in-sample error E_{in} and out-of-sample error E_{out}, respectively, and set h, as the hypothesis that generated v and seeks to approximate the unknown target function f. However, the learning process involves putting into trial many hypotheses, usually infinite of them, to find the best hypothesis g. Because the selection of g depends on a training set and occurs after evaluating the performance of a set or sets of hypotheses, the key assumption of randomness that held the Hoeffding Inequality is violated (Abu-Mostafa et al., 2012, p. 22-23). To show this violation, the results of a simulation using coin flips is shown in **Figure 1**.

One thousand trials were ran, and at each iteration, 1,000 coins were each flipped 10 times, and the proportion of heads for a) the first coin, b) a coin selected at random, and c) that corresponding to the coin with lowest proportion was recorded, along with their distributions. After running all the trials, both sides of the Hoeffding Inequality were evaluated for each of the 3 set of results.[1] As expected, the bound holds for the first coin and the coin chosen at random, but not for the coins with minimum frequency. The histograms show that the distribution of a non-random v does not approximate the target distribution, in this case, a binomial distribution. Translated into a learning context, selecting the hypothesis that minimizes $E_{in}(g)$, instead of $E_{in}(h)$ is akin to choosing the coin exhibiting the lowest proportion and does not necessarily reflect the error distribution of the target E_{out}. Perhaps, such discrepancy is a manifestation of overfitting.

Figure 1.

oeffding Inequality Simulation: Effect of Random vs Non−Random Hypothesis Selection



In order to obtain a realistic bound, the fact that $E_{in}(g)$ is selected instead of $E_{in}(h)$ in a data dependent way, has to be accounted for, in the same way adjustements are made for large-scale multiple testing (Abu-Mostafa et al., 2012, p.23). One way is to use the union bound:

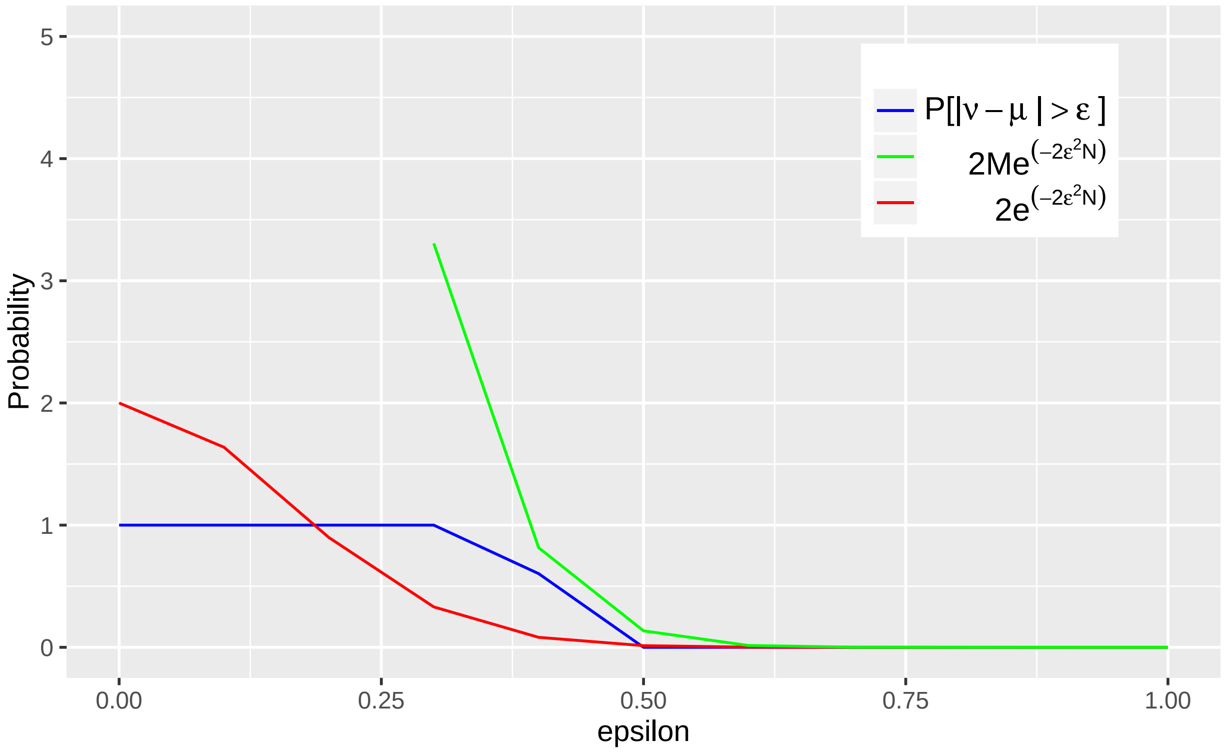
$P[|v-u| > \epsilon] \leq 2Me^{-2\epsilon^2 N}$, for any $\epsilon > 0$,

where M corresponds to the number of hypotheses tested, which must be finite. The union bound states that the probability of any one event belonging to a set of events is less than the sum of the probabilies of all events in the set (Abu-Mostafa et al., 2012, p.41). In the coin simulation, although $M = 1000$, for illustration purposes, the union bound has been plotted with $M = 10$.[2] As observed in **Figure 2**, the bound holds, but drastically overestimates it.

[1] Based on Excercise 1.10 (Abu-Mostafa et al., 2012, p.23).

[2] This simulation was not part of the original excerise.

Figure 2.
Minimum Proportion Coin with Union Bound



2.2 Applying the VC Inequality to evaluate learning with an infinite hypotheses set

Unlike the scenario depicted in **Figure 1**, where the Hoeffding Inequality was used for a single hypothesis and the union bound for a finite hypothesis set, a typical machine learning situation involves infinite hypothesis sets. The key insight is to observe that the majority of the infinite number of possible hypotheses overlap to generate the same learning outcome and thus, the trick is to generate a reliable bound that accounts for such overlap by characterizing an effective, finite number of hypotheses, m_H, based on the given data set, instead of the entire input space (Abu-Mostafa et al., 2012, p.41-42). Such characterization is brought about by the VC Dimension, d_{VC}, which helps set a polynomial bound to the effective number of hypotheses, $m_H \leq N^{d_{VC}} + 1$ (Abu-Mostafa et al., 2012, p.50). Proving that the bound is polynomial was a breakthrough and, indeed, is what makes learning feasible, as long as there is enough N, relative to the VC Dimension (Abu-Mostafa et al., 2012, p.53). Furthermore, it has been shown that the VC Dimension can be conceptualized to be the effective number of parameters of a learning model (Abu-Mostafa et al., 2012, p.52). For example, the VC Dimension for linear regression is $P + 1$, where P corresponds to the number of parameters in the linear model (Abu-Mostafa et al., 2012, p.78).

The VC Dimension is useful because it can be used to substitute the union bound's M in the Hoeffding Inequality (Abu-Mostafa et al., 2012, p.53). With some adjustements, a new bound, the VC Inequality, can be formulated as,

$P[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 4m_H(2N)e^{\frac{-1}{8}\epsilon^2 N}$ **(Equation 1)**

It can be re-written to express the VC Bound, as the discrepancy between $E_{in}(g)$ and $E_{out}(g)$, by selecting a tolerance level δ (how often the bound is violated) and stating with probability at least $1 - \delta$ that,

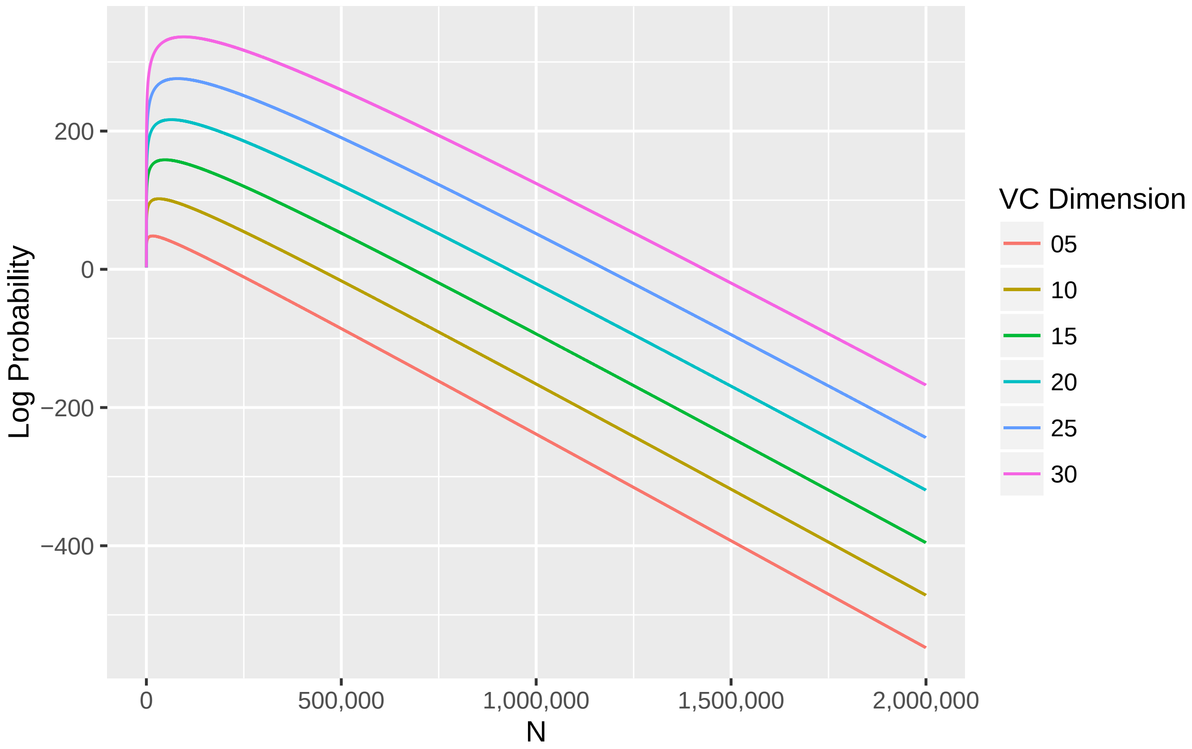
$E_{out}(g) - E_{in}(g) \leq \sqrt{\frac{8}{N} ln \frac{4m_H(2N)}{\delta}}$

For simplicity, $m_H(2N)$ can be replaced by its upper bound,

$E_{out}(g) - E_{in}(g) \leq \sqrt{\frac{8}{N} ln \frac{4((2N)^{d_{VC}}+1)}{\delta}}$ **(Equation 2)**

As observed, a logarithmically growing m_H cannot outdo a negative expotential, and thus, as mentioned earlier, there is a theoretical guarantee that the generalization error will be small with large N (Abu-Mostafa et al., 2012, p.51). To see this, a simulation of the bound as per Equation 1 was performed using different values of VC Dimension and N, with ϵ fixed to 0.05.[3] The results are presented in **Figure 3** below.

Figure 3.
VC Inequality Sensitivity to VC Dimension and Sample Size



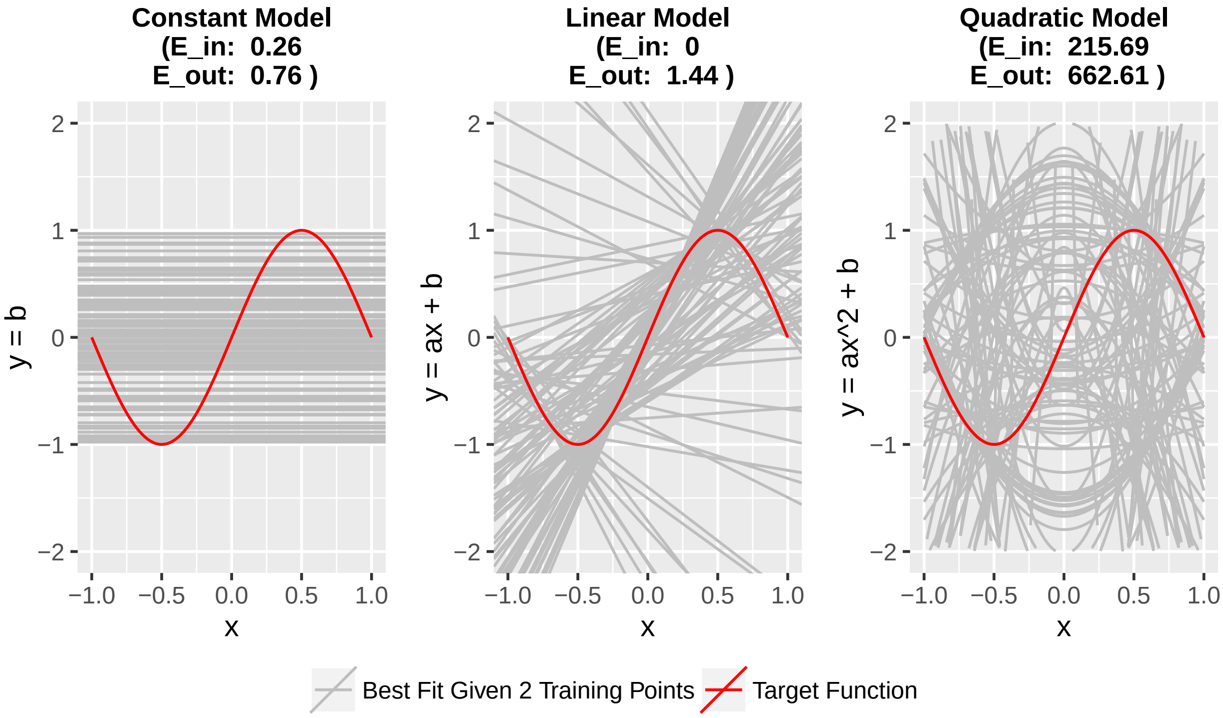
In summary, VC Theory informs us that provided a data set of size N and our choice of a learning algorithm, whether simple or complex, we can obtain a sense on the expected level of generalization. For example, in most learning situations, a hypothesis of the form $h(x) = ax + b$ will perform better than one of the form $h(x) = b$, provided there are is sufficient N. However, provided with only 2 training points, a linear regression model can prove to be too sophisticated than one based on a constant, as the below simulation shows.[4] In fact, it is so sophisticated that it will perfectly overfit with an $E_{in} = 0$, at the expense of a larger E_{out}, compared to the simple constant model. This is simulated in **Figure 4.**

The most important lesson taught by VC Theory, as espoused by Abu-Mustafa et al. (2012), is to not haphazardly apply the most sophisticated learning algorithm in the toolkit, but to match data quantity and quality with model complexity. This point is further explored in the following section.

[3] Replication of a plot presented in a lecture from the online course "Learning From Data" offered by Caltech. The image correspondes to slide 21, downloadable from http://bit.ly/28M0PaA.

[4] Based on Homework 4, Problem Set 7 from the online course "Learning From Data" offered by Caltech, downloadable from http://bit.ly/28M1Dw2 .

Figure 4.
Effect of Learning 2 Data Points with Increasingly Complex Linear Models



3. Conceptual Applications for Obtaining Good Generalization Based on VC Theory

As a conceptual tool, VC Theory is powerful. Although quantifying the VC Dimension and computing the VC Bound for every data science project is not done in practice, and perhaps not possible without the use of loose heuristic methods, VC Theory can help guide decision making and create a sense of whether a learning problem is in fact learnable.

For this section, when applicable, the time series forecasting competition that was held in our Financial Econometrics class will be used as a caste study in order to show some practical applications of VC Theory. Furthermore, for simplicity, the learning model to be used in the following examples is linear regression, which as mentioned above has a VC Dimension of '# of parameters + 1'. The dataset for the competition consisted of 500 observations and a mixture of 50 relevant and non-relevant variables.

a) Estimating the required size of the data to make training feasible

The VC Inequality form stated in Equation 2 can be rephrased in terms of N, where we express the the left hand side of the inequality, $E_{out}(g) - E_{in}(g)$ as the allowed generalization error, ϵ,

$\epsilon \leq \sqrt{\frac{8}{N} ln \frac{4((2N)^{d_{VC}}+1)}{\delta}} \rightarrow N \leq \frac{8}{\epsilon^2} ln(\frac{4((2N)^{d_{VC}}+1)}{\delta})$ **(Equation 3)**

If we wish to obtain a result with 95% confidence that the generalization error is at most 0.05 using only 1 variable, the required sample size as per Equation 4 is approximately $N = 92,000$. However, because the VC Bound is distribution free, the sample size requirement is grossly overestimated and for practical purposes, a rule of thumb of $N \geq 10d_{VC}$ is recommended.[5] If so, even if the full 50 variables are used, the required N for using a linear model would be greater than the given set of 500 data points. No wonder, data science students who employed more complex learning models, such as Random Forest, which would imply a larger VC Dimension, were not able to beat the winning team's linear regression.

b) Estimating expected VC Bound

Conversely, as shown in **Figure 3**, a bound can be estimated given N as per **Equation 2**. Once again, at N fixed at 500, a 95% confidence requirement, and using 1 variable, the generalization error is bounded by 0.54. This means, in the worst case of overfitting with an $E_{in}(g) = 0$, the probability that the out-of-sample prediction is wrong is bounded by a coin-flip. However, with a more complex model, perfect overfit would be accompanied with a larger bound. Once again, this shows why linear regression may have outperformed fancier ensemble models.

c) Understanding regularization

Regularization is a technique to reduce overfitting by controlling model complexity, such as the size of the coefficients (ridge regression) or number of dimensions of a linear model (LASSO). From a bias-variance trade-off perspective, it implies reducing out-of-sample variance by sacrificing some bias, with the expectation that their combined effect results in a lower E_{out} (Abu-Mostafa et al., 2012, p. 128).

In order to understand regularization using the VC framework, we refer back to Equation 2 and notice that the right hand side of the inequality is a function of N, d_{VC}, and δ. Therefore, the bound can be reduced by using a hypothesis set with low VC Dimension, although should not be overdone at the expense of the model's ability to approximate the target function (underfitting). However, unliked hard regularizers like LASSO which force coefficients to zero, many soft regularization techniques, such as ridge regression, preserve a model's VC Dimension, while still improving generalization. This favorable result may not fully be explained by VC Theory, but implicitly, regularization seems to achieve the equivalence of a reduction of the VC Dimension by limiting the size of the parameter space, and in turn, the size of the hypothesis set (Abu-Mostafa et al., 2012, p. 137).

d) Understanding validation

Splitting the data into a training and test set has an intuitive appeal, to the extent that it is practiced without knowledge of its full theoretical foundation. Its purpose is to estimate E_{out} using the test set error, instead of the training set error, E_{in} (Abu-Mostafa et al., 2012, p. 138). The estimate is reliable because the test set is not biased and such approximation is possible because the test set is only used to evaluate the out-of-sample error of the final best hypothesis, g, obtained during training. Therefore, the effective number of hypotheses equals 1 and thus, instead of relying on the VC Inequality that accounts for the existence of an infinite number of hypothesis, one can rely on the tighter bounded Hoeffding Inequality (Abu-Mostafa et al., 2012, p. 139).

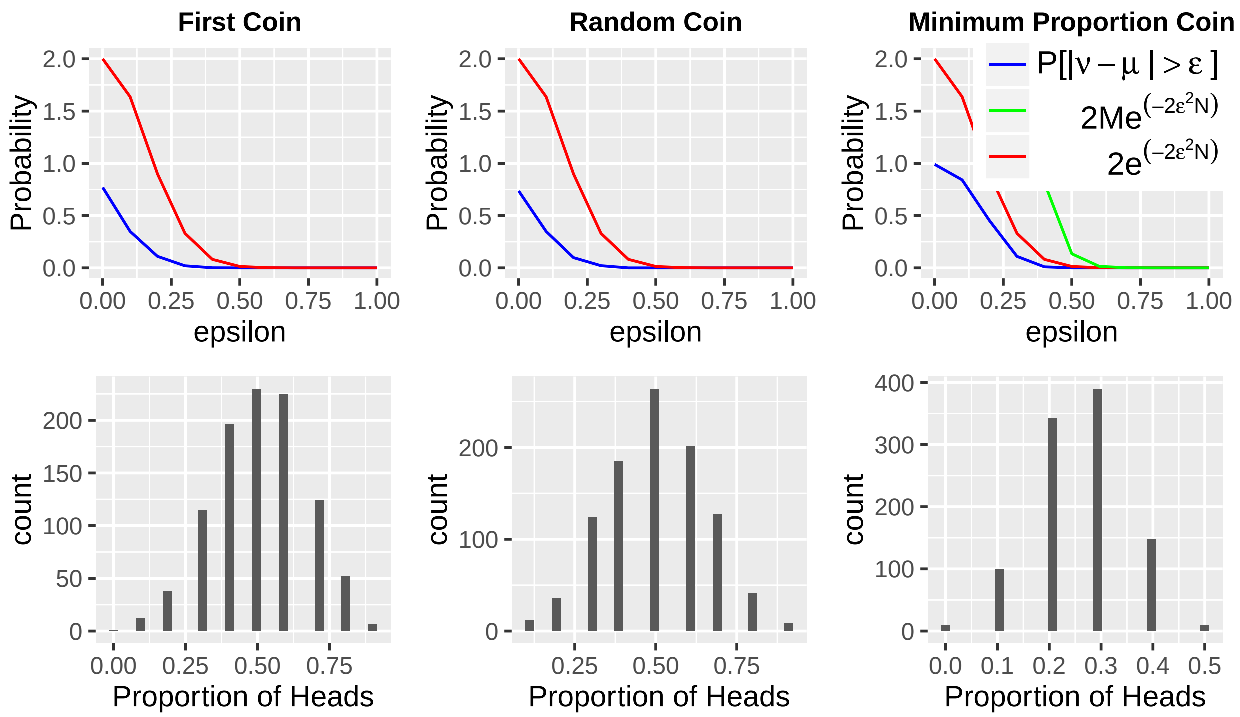
In practice, a couple of final best hypotheses would be tested on the test set for the purpose of selecting the best value out of a finite set of candidate regularization parameters or the final model out of a finite set composed by the best in-class performing logistic regression, SVM, and random forest model, to name some examples. Therefore, in the context of validation, the union bound would be appropriate (Abu-Mostafa et al., 2012, p. 143). If we redo the first coin simulation above with 10 coins[6], instead of the original 1,000, we notice two things: a) Hoeffding Inequality holds and b) the distribution of heads does not deviate significantly from the target distribution, albeit at the cost of a small optimistic bias.

[5] Lecture from the online course "Learning From Data" offered by Caltech: slide 21, downloadable from http://bit.ly/28M0PaA .

[6] Not part of the original textbook exercise.

Figure 5.

Hoeffding Inequality 10−Coin Simulation: Random vs Non−Random Hypothesis Selectio



References

1. Abu-Mostafa, Y. S., Magdon-Ismail, M., & Lin, H. (2012). Learning from data: a short course. [United States]: AMLBook.com.