Machine Learning and the String Landscape

Brent D. Nelson

String Pheno 2017, Virginia Tech

hep-th/1607.00655, with J. Carifio, J. Halverson, and D. Krioukov

Northeastern University

NU HET Data Science: R. Altman, W. Cunningham, C.Long, T. Eliassi-Rad

When is Data ”Big Data”?

⇒ Traditional definition: data is ”Big” when it has one or more of the following properties:

- High volume
- High velocity
- High variability

When is Data "Big Data"?

⇒ Traditional definition: data is "Big" when it has one or more of the following properties:

- High volume: landscape studies tend to involve $\mathcal{O}(10^9)$ or more explicit objects
- High velocity
- High variability: data comes in form of binary objects, integers, arrays and tensors, and size of objects can vary with internal parameters (e.g. h^{11})

⇒ Data sets in string phenomenology satisfy at least two of three criteria

The Biggest Big Data

$\Rightarrow$ The string landscape: the biggest Big Data imagineable?

$\mathcal{O}(100)$ **GB** Storage capacity of my IPhone

$\mathcal{O}(1)$ **TB** Storage capacity of my desktop computer

$\mathcal{O}(10)$ **TB** A large sized research library

$\mathcal{O}(100)$ **TB** All credit card transactions in U.S., per year

$\mathcal{O}(10)$ **PB** Annual data generated by an LHC detector

$\mathcal{O}(100)$ **PB** Typical data storage on private could services

$\mathcal{O}(10^3)$ **PB** *Estimated* data stored at NSA Utah Data Center

$\mathcal{O}(10^6)$ **PB** *Estimated* annual volume of all internet traffic

The Biggest Big Data

$\Rightarrow$ The string landscape: the biggest Big Data imagineable?

$\leftarrow$ "Unpacked" Toric CY3 database (rossealtman.com), $h^{1,1} \leq 6$

$\mathcal{O}(100)$ **GB** Storage capacity of my IPhone

$\leftarrow$ Kreuzer-Skarke 4D polytope data, raw

$\mathcal{O}(1)$ **TB** Storage capacity of my desktop computer

$\mathcal{O}(10)$ **TB** A large sized research library

$\mathcal{O}(100)$ **TB** All credit card transactions in U.S., per year

$\mathcal{O}(10)$ **PB** Annual data generated by an LHC detector

$\mathcal{O}(100)$ **PB** Typical data storage on private could services

$\leftarrow$ *Estimate* of set of all FRSTs of KS 3D polytope set (lower bound)

$\mathcal{O}(10^3)$ **PB** *Estimated* data stored at NSA Utah Data Center

$\mathcal{O}(10^6)$ **PB** *Estimated* annual volume of all internet traffic

$\leftarrow$ *Estimate* of size of all CY3 data implicitly contained in KS 4D polytope set

Recent Interest in the Field

⇒ Machine Learning (Neural Networks) and String Theory

1706.02714 Yang-Hui He

1706.08503 Krefl and Seong

1706.07024 Fabian Ruehle, Friday at noon

⇒ Computational Complexity in the Landscape

1706.06430 Denef, Douglas, Greene, Zukowski

1706.08503 Bao, Bousso, Jordan, Lackey

The Goals of Machine Learning

Deep Data Dive Via training a model on a subset of an ensemble, it is sometimes feasible to make high accuracy feature predictions that are much faster than conventional techniques, allowing for far greater exploration of the dataset.

Conjecture Generation The decision function of a trained model may naturally lead to a sharp conjecture that can be rigorously proven.

Feature Identification and Extraction When input data is of high dimensionality, or exhibits redundant information, models can identify those properties (*features*) of the data that are most correlated with desired outcomes. This is often one of the primary goals of landscape surveys in string theory.

What is Machine Learning, Anyway?

⇒ Machine learning is a set of algorithms that train on a data set in order to make predictions on unseen data

- The output from most machine learning algorithms is a function, commonly called a *model*
- Model takes a specified set of inputs and produces a unique output value for the characteristic in question

⇒ Supervised machine learning

- Training step is performed by allowing the algorithm to experience a number of input $\rightarrow$ output pairs
- Most common (and generally most successful) form of machine learning
- Makes sense in string theory contexts where the whole point of laboriously computing the dataset is to extract particularly interesting physical quantities

Classification vs. Regression

⇒ Two broad classes of problems to attack with machine learning techniques

⇒ Most machine learning techniques can be operated in either “mode”

Classification

- Can be done supervised or unsupervised (clustering)
- Use input data to assign classes (labels) to the objects
- Goal can be a simple binary question
- Model is evaluated on the basis of its *accuracy*

Regression

- Generally applicable to supervised learning only
- Use input data to predict a continuous, real-valued output
- Both linear and non-linear methods can be employed
- Model is evaluated on the basis of *statistical variance*

Model Training and Validation

⇒ How do we evaluate the best approach to a problem?
How do we know a model is 'working'?

- Simplest method: designate a train-test split
- Example: train on 75% of the input $\rightarrow$ output pairs, try to predict (known) outputs on remaining 25%

⇒ A better approach is k*-fold cross-validation*

- Data is divided into k equal subsets
- Each model trained k separate times, in each reserving only one of the k partitions for testing, and using the other $k-1$ folds for training
- Minimizes training-sample bias, maximizes exposure to 'known' data

⇒ For regression tasks, scoring metric will be the mean absolute percent error

$$\text{MAPE} := \frac{100}{N} \times \sum_{i=1}^{N} \left| \frac{A_i - P_i}{A_i} \right|,$$

where P_i and A_i are the predicted and actual values for the output quantity

Techniques Employed

- Linear Regression (LIR)
- Logistic Regression (LR)
- k-Nearest Neighbors (KNN)
- Classification and Regression Tree (CART)
- (Gaussian) Naive Bayes (NB)
- Linear Discriminant Analysis (LDA)
- Support Vector Machines (SVM)

Deep Data Dive: Counting Smooth F-Theory Compactifications

Description of the Problem

⇒ GOAL: estimate the number of fine regular star triangulations (FRST) of all 4319 three-dimensional reflexive polytopes

- Each such FRST determines a smooth weak-Fano toric variety.
- Give rise to smooth F-theory compactifications without non-Higgsable clusters

⇒ METHOD: Estimate the number of FRSTs of the 3d polytope via product of the number of FRTs of its codimension one faces (i.e. the facets)

- Procedure shown to work reasonable well by brute force computation up to $h^{1,1}(B) \leq 22$
- Seek to train a model (A) to predict number of FRTs (n_T) given gross properties of the facet (F)

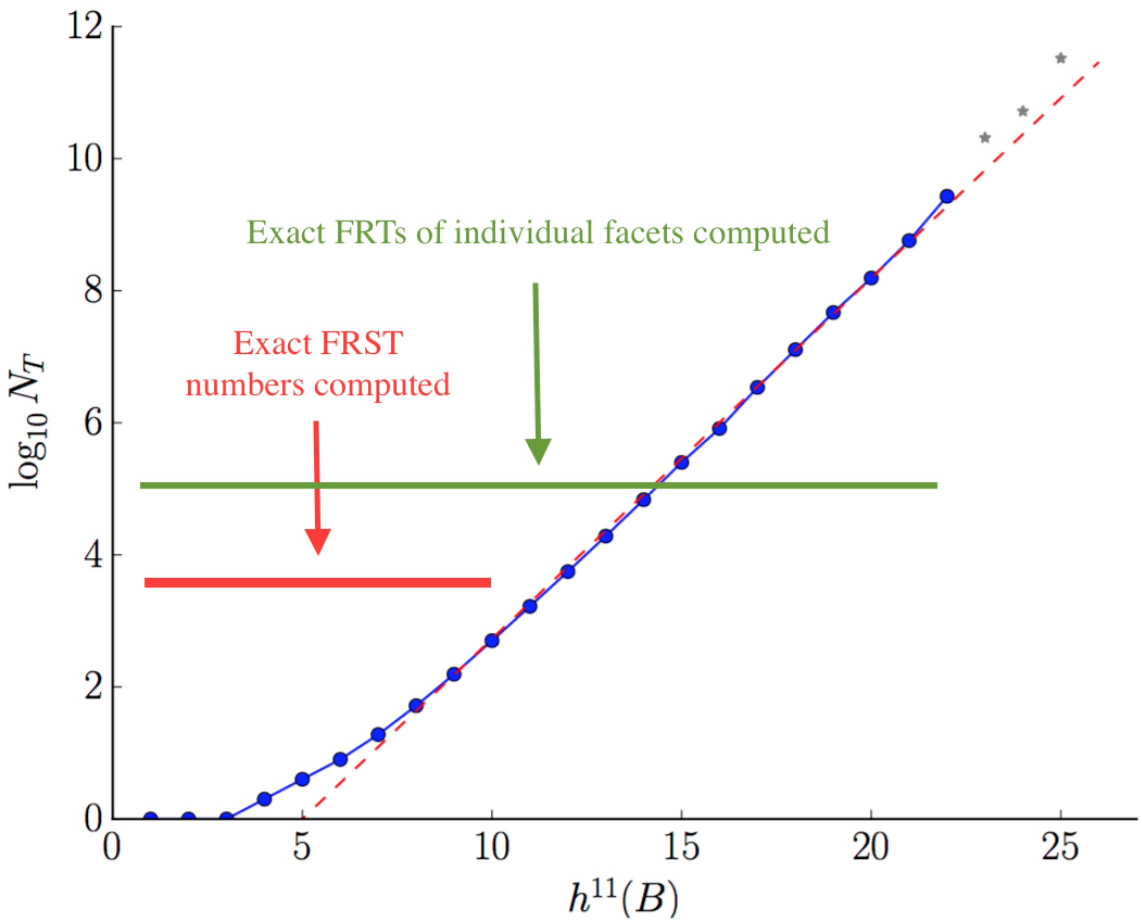
$$F \longrightarrow (n_p, n_i, n_b, n_v) \xrightarrow{A} n_T,$$

- ⋆ n_P = number of points
- ⋆ n_i = number of interior points
- ⋆ n_b = number of boundary points
- ⋆ n_v = number of vertices

Setting Up the Training

⇒ Subtlety: how best to set up appropriate training/validation data?

- Direct computation of FRST performed for cases with $h^{1,1}(B) \leq 10$
- Direct computation of FRTs of the facets to $h^{1,1}(B) \leq 22$



- Growth in FRSTs changes at $h^{1,1} = 7$, approaching linear in log plot
- For extrapolation to the highest $h^{1,1}$, how much should the machine be allowed to 'train' on the low $h^{1,1}$ data?
- Optimize the learning by choosing a variety of $h^{1,1}_{\text{min}}(B) \leq h^{1,1}_{\text{train}}(B) \leq h^{1,1}_{\text{max}}(B)$ regimes:

$$h^{1,1}(B)_{\text{min}} \in \{1 - 10\}$$

$$h^{1,1}(B)_{\text{max}} \in \{14 - 18\}$$

Model Selection, Validation, and Refinement

⇒ Perform ten-fold cross-validation to train four models, using MAPE as evaluation metric, for all combinations of $\{h^{1,1}_{\rm min}(B), h^{1,1}_{\rm max}(B)\}$

- Models considered for this problem: LDA, KNN, CART, NB

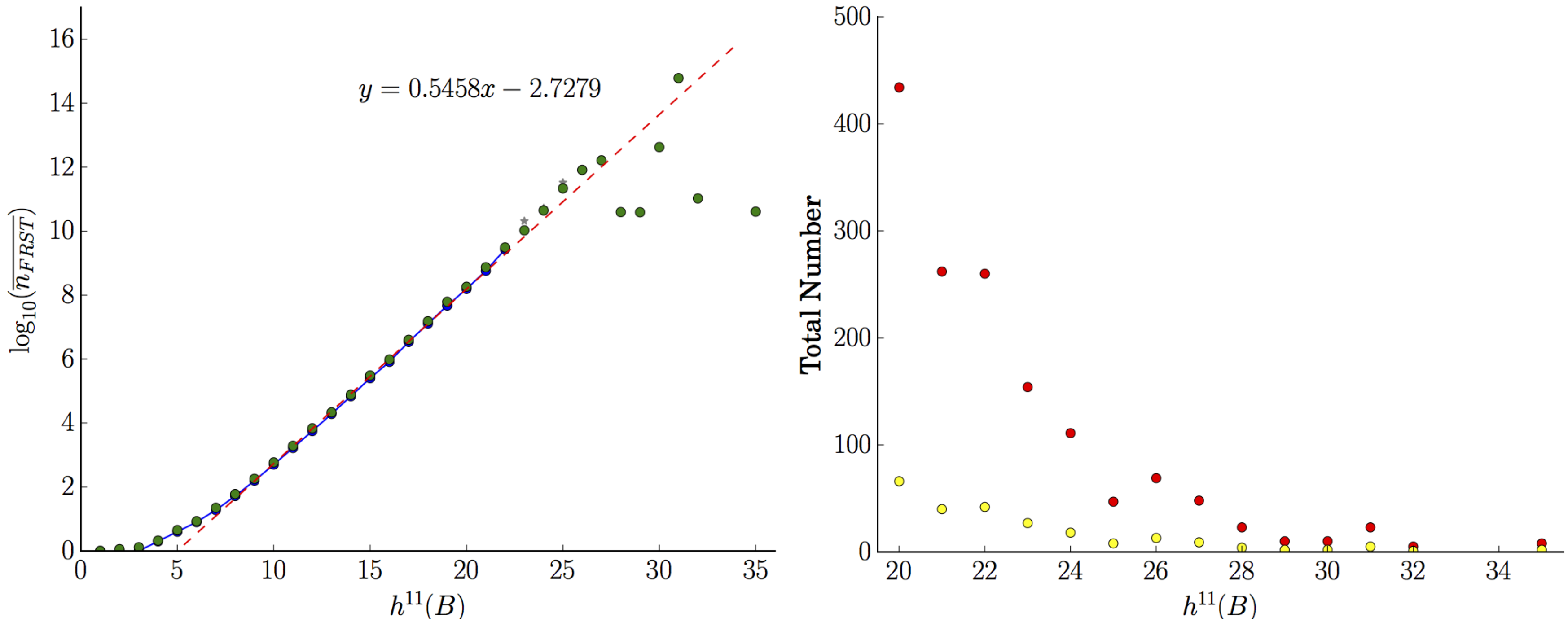
- Evaluate the MAPE on the test/train data, as well as $h^{1,1} = 19,\ 20,\ 21$ extrapolation training data

⇒ Best combination of MAPE on train/test data and extrapolation data was regression tree (CART) method for $\{h^{1,1}_{\rm min}(B), h^{1,1}_{\rm max}(B)\} = \{4, 18\}$

- MAPE for train/test data: 5.5%; MAPE for $h^{1,1}(B) = 19, 21$: 11.2%, 17.4%

⇒ Now re-train the CART algorithm on full set from $\{h^{1,1}_{\rm min}(B), h^{1,1}_{\rm max}(B)\} = \{4, 21\}$

Results



$\Rightarrow$ Good predictions for $22 \leq h^{1,1}(B) \leq 27$, less reliable thereafter

- Remaining cases in the tail of the distribution across entire 4319 polytopes (more description of these cases!)
- Assuming a linear extrapolation, one predicts $n_{FRST} \sim O(10^{15}) - O(10^{16})$ at $h^{11}(B) = 35$

$\Rightarrow$ Consistent with the predicted bounds $5.8 \times 10^{14} \lesssim n_{FRST} \lesssim 1.8 \times 10^{17}$ from Halverson, Tian (2016)

Conjecture Generation: Rank($\mathcal{G}$) in F-Theory Ensembles

Background: What Determines the Gauge Group?

$\Rightarrow$ $D7$-branes localized at discriminant locus, defined by $\Delta = 4f^3 + 27g^2 = 0$

- The order of vanishing of the polynomials f and g along some $x_i = 0$ determines the gauge group on $x_i = 0$
- Construction of the most general polynomials f and g appearing in Δ assisted by computation of auxiliary polyhedra $\Delta_f = \{m \in \mathbb{Z}^3 \,|\, m \cdot v_i + 4 \geq 0 \ \forall i\}$ and $\Delta_g = \{m \in \mathbb{Z}^3 \,|\, m \cdot v_i + 6 \geq 0 \ \forall i\}$
- Allowed monomials in f and g correspond to integral points in Δ_f and Δ_g

$\Rightarrow$ Adornment of polytope with edge and face trees represents a topological transition $B \to B'$

- Each new leaf introduces new rays v_i into the definition of Δ_f and Δ_g
- This may result in the removal of certain monomials $m_f \in \Delta_f$ or $m_g \in \Delta_g$

$\Rightarrow$ This transformation $\Delta_f,\, \Delta_g \to \Delta'_f,\, \Delta'_g$ changes the final gauge group in the theory

- Thus, the existence of a gauge group at some particular locus in base B' is dependent on the tree structure everywhere over the original base B

Machine Learning the Value of $rk(G)$

$\Rightarrow$ GOAL: predict the rank of the geometric gauge group in the large ensemble of F-theory geometries

- Assumption: the overall rank is determined by the number of leaves of various heights above the base
- Let H_i be the number of height i leaves in B
- We seek to train a model A to predict the rank of the resulting gauge group $rk(G)$ on the base B

$$B \longrightarrow (H_1, H_2, H_3, H_4, H_5, H_6) \xrightarrow{A} rk(G)$$

$\Rightarrow$ We perform a 10-fold cross validation with sample size 1000 and algorithms LR, LIR, LDA, KNN, CART, NB, SVM

- The linear regression gave the best results, having MAPE 0.013
- The decision function is

$$\begin{aligned} rk(G) &= 302.54 - 1.1102 \times 10^{-16}\, H_1 + 3.9996\, H_2 + 1.9989\, H_3 \\ &\quad + 1.0007\, H_4 + 1.3601 \times 10^{-3}\, H_5 + 1.1761 \times 10^{-3}\, H_6 \end{aligned}$$

Connection to a Known Result

$\Rightarrow$ Height 1 leaves are facet interior points that are always present, so $H_1 = 38$

- Thus the H_1 term is effectively a constant, with $H_1 = 38$
- Using the fact that $304 = 38 \times 8$, one can rewrite the regression equation equivalently as

$$\begin{aligned} rk(G) = \quad & -1.46 + 8\,H_1 + 3.9996\,H_2 + 1.9989\,H_3 + 1.0007\,H_4 \\ & +1.3601 \times 10^{-3}\,H_5 + 1.1761 \times 10^{-3}\,H_6 \end{aligned}$$

$\Rightarrow$ Since any leaf can only contribute an integer of rank to the gauge group, we can round the coefficients to the nearest integer

$\Rightarrow$ We also expect the intercept to vanish, and it is indeed small (-1.46) relative to an expected rank of $\mathcal{O}(2000)$

$\Rightarrow$ With these considerations taken into account, we have a prediction:

$$rk(G) \simeq 8\,H_1 + 4\,H_2 + 2\,H_3 + \;H_4$$

Genesis of a Formal Conjecture

⇒ The relationship between the rank of the gauge group $rk(G)$, and the heights H_i suggests the following conjecture

Conjecture: with high probability, height 1 leaves have gauge group E_8, height 2 leaves have gauge group F_4, height 3 leaves have gauge group G_2 or A_2, and height 2 leaves have gauge group A_1.

⇒ The above can be made more rigorous given the construction algorithm described previously:

Refined Conjecture: Let v be a leaf $v = av_1 + bv_2 + cv_3$ built on roots $v_{1,2,3}$ whose associated divisors carry E_8. Then if the leaf has height $h_v = 2, 3, 4$ its associated gauge groups are F_4, $\in \{G_2, A_2\}$, and A_1, respectively.

⇒ Ultimately, this leads to an even more precise theorem, which can be proven:

Theorem: Let v be a leaf $v = av_1 + bv_2 + cv_3$ with v_i simplex vertices in F. If the associated divisors $D_{1,2,3}$ carry a non-Higgsable E_8 seven-brane, and if v has height $h_v = 1, 2, 3, 4, 5, 6$ it also has Kodaira fiber $F_v = II^*, IV^*_{ns}, I^*_{0,ns}, IV_{ns}, II, -$ and gauge group $G_v = E_8, F_4, G_2, SU(2), -, -$, respectively.

Conjecture Generation: How it Should Work

1. ***Variable Selection.*** Based on knowledge of the data, choose input variables X_i that are likely to determine some desired output variable Y. In the example, this was recognizing that $X_i = H_i$ may correlate strongly with gauge group.

2. ***Machine Learning.*** Via machine learning, train a model to predict Y given X_i with high probability. In this example, a 10-fold cross validation was performed, and it was noted that the highest accuracy came from a linear regression.

3. ***Conjecture Formulation.*** Based on how the decision function uses X_i to determine Y, formulate a first version of the conjecture. In this example, the first version of the conjecture arose naturally from the linear regression and basic dataset knowledge.

4. ***Conjecture Refinement.*** The original conjecture arose from a model that was trained on a dataset that is subject to sampling assumptions. Those assumptions may lead to high probability properties critical to proving the conjecture; refine accordingly based on them. In the example, we used the high frequency of E_8 on the ground.

5. ***Proof.*** After iterating enough times that the conjecture is precise and natural calculations or proof steps are obvious, attempt to prove the conjecture.

Conjecture Generation: Frequency of E_6 Sectors in F-Theory Ensembles

The Puzzle of E_6

⇒ As we saw above, the simple factors in the generically semi-simple group G were $G_i \in \{E_8, F_4, G_2, A_1\}$

- These groups only have self-conjugate representations
- None particularly suggestive of the Standard Model

⇒ E_6 (and $SU(3)$) also exists: random sampling suggests probability $\simeq 1/1000$

- Conditions under which E_6 or $SU(3)$ existed were not known
- Sampling revealed E_6 only arose on a particular distinguished vertex $v_{E_6} = (1, -1, -1)$
- Training models on tree height data alone did not produce high accuracy results – better variable selection required

⇒ GOAL: determine whether a particular leaf supports a particular gauge group

⇒ If we focus exclusively on E_6, this is a *binary classification problem*.

Finding Better Input Variables

$\Rightarrow$ The gauge group on low lying leaves depends on the heights of trees placed at various positions around the polytope

$\Rightarrow$ The task is to determine those new leaves v that could cause a particular m_f or m_g to be chopped off of Δ_f or Δ_g

- Consider the 'headroom' associated with each of integral points
- Construct the set $S_{a,v_1} := \{v \in V | v = av_1 + bv_2 + cv_3, \;\; a, b, c \geq 0\}$
- The size of this set depends on the choice of v_1 and the value of a (limited by headroom)

$\Rightarrow$ For each v in the polytope, find the value $a_{\max}$ such that $S_{a>a_{max},v_1}$ is empty, and record a_{max} and $|S_{a_{max},v_1}|$

$\Rightarrow$ Now ready to train models to determine if E_6 is present on v_{E_6}, with accuracy being the evaluation metric

$$\Delta_1^\circ \longrightarrow (a_{max}, |S_{a_{max},v}|) \quad \forall v \in \Delta_1^\circ \qquad \xrightarrow{A} \qquad E_6 \text{ on } v_{E_6} \text{ or not,}$$

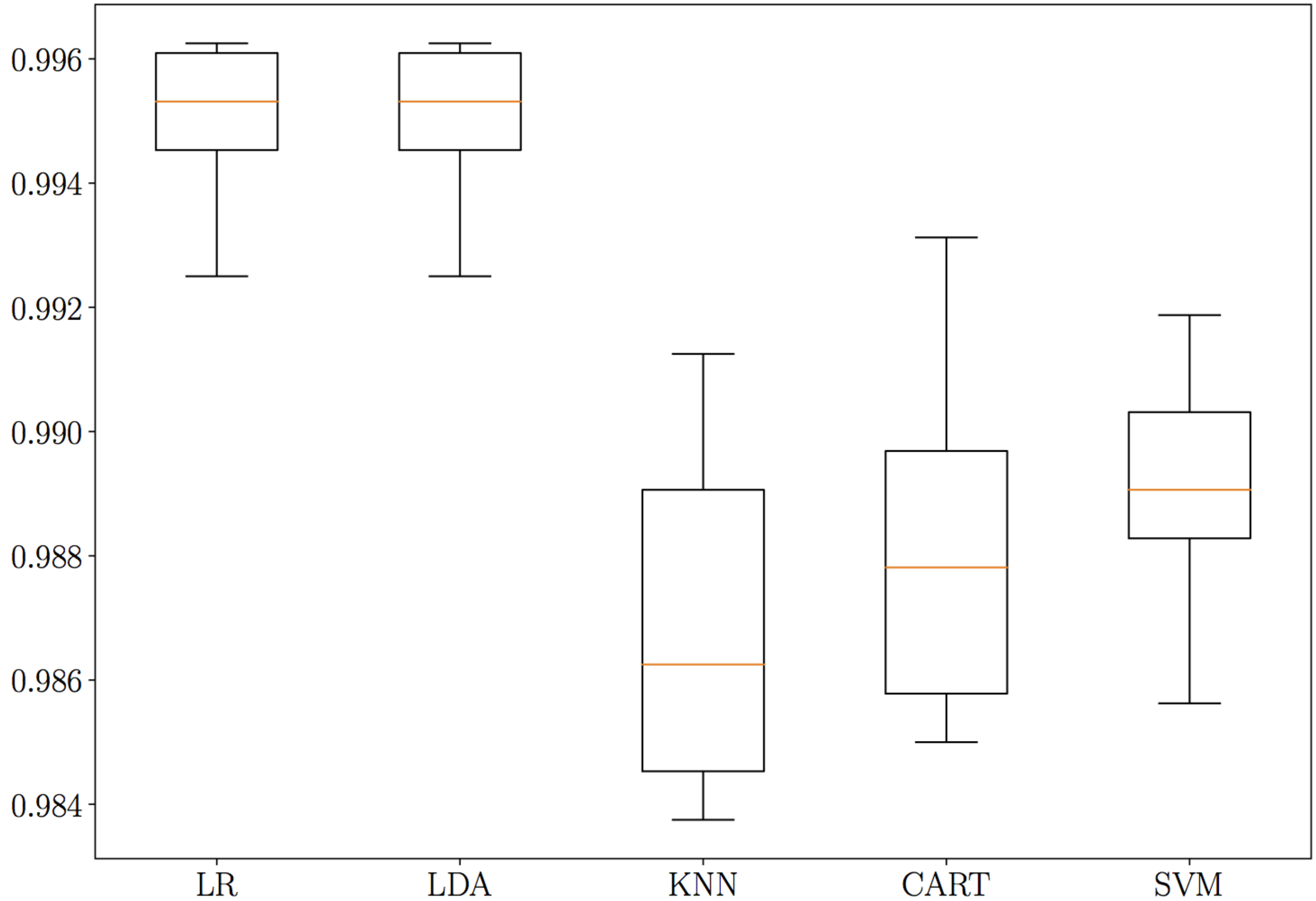
Model Training

⇒ We train on 20000 samples, but it is important to change the sampling assumptions slightly

- Under the previous assumptions, ~ 20 of the samples would have E_6 on v_{E6} and the rest would not
- Trained models might tend to predict no E_6 uniformly, leading to an "accuracy" of $.999$.
- We will therefore train the model on an enriched sample: 10000 examples with E_6 and 10000 without

⇒ Perform 10-fold cross validation using algorithms LR, LDA, KNN, CART, SVM

Model Results



	LR	LDA	KNN	CART	SVM
$50/50$ Validation Set	$.994$	$.994$	$.982$	$.987$	$.989$
Unenriched Set	$.988$	$.988$	$.981$	$.988$	$.983$.

Example of Factor Analysis

⇒ Can we see what is driving the model? Yes!

- Perform a factor analysis – try to reduce the dimensionality of the $2 \times 38 = 76$ integers fed into the model
- The analysis reveals a key feature – the most important data is the input pair where $v_1 = v_{E_6}$ itself!

⇒ How much does this one pair of numbers capture of the whole model? Train again, this time only utilizing $(a_{max}, |S_{a_{max}, v_{E6}}|)$

	LR	LDA	KNN	CART	SVM
$50/50$ Validation Set	$.994$	$.994$	$.994$	$.994$	$.994$
Unenriched Set	$.988$	$.988$	$.988$	$.988$	$.983$.

Conjecture Generation

⇒ Looking in even more detail, the logistic regression model provides hints to an organizing principle

a_{max}	$\lvert S_{a_{max},v_{E6}} \rvert$	Pred. for E_6 on v_{E_6}	Hyperplane Distance
4	5	No	0.88
4	6	No	0.29
4	7	Yes	-0.31
4	8	Yes	-0.90
4	9	Yes	-1.50
$\ldots$	$\ldots$	$\ldots$	$\ldots$
4	21	Yes	-8.64
4	22	Yes	-9.23
4	23	Yes	-9.83
4	24	Yes	-10.42
5	1	No	7.34
5	2	No	6.75
$\ldots$	$\ldots$	$\ldots$	$\ldots$
5	8	No	3.18
5	9	No	2.59
5	10	No	1.99
5	11	No	1.40
5	12	No	0.80

Conjecture Verification

Initial Conjecture: If $a_{max} = 5$ for v_{E6}, then v_{E6} does not carry E_6. If $a_{max} = 4$ for v_{E6} it may or may not carry E_6, though it is more likely that it does.

$\Rightarrow$ After some thinking (see the manuscript), one arrives at the following Theorem, which can be proven

Theorem: Suppose that with high probability the group G on v_{E_6} is $G \in \{E_6, E_7, E_8\}$ and that E_6 may only arise with $\tilde{m} = (-2, 0, 0)$. Given these assumptions, there are three cases that determine whether or not G is E_6.

a) If $a_{max} \geq 5$, $\tilde{m}$ cannot exist in Δ_g and the group on v_{E_6} is above E_6.

b) Consider $a_{max} = 4$. Let $v_i = a_i v_{E_6} + b_i v_2 + c_i v_3$ be a leaf built above v_{E_6}, and $B = \tilde{m} \cdot v_2$ and $C = \tilde{m} \cdot v_3$. Then G is E_6 if and only if $(B, b_i) > 0$ or $(C, c_i) > 0$ $\forall i$. Depending on the case, G may or may not be E_6.

c) If $a_{max} \leq 3$, $\tilde{m} \in \Delta_g$ and the group is E_6.

Conjecture Verification

⇒ The theorem produces a sharp prediction for the probability of E_6 across the whole dataset

$$P(E_6 \,\text{on}\, v_{E_6} \,\text{in}\, T) = \left(1 - \frac{36}{82}\right)^9 \left(1 - \frac{18}{82}\right)^9 \simeq .00059128$$

⇒ We can check if this is reasonable by simply counting the number of E_6 instances across five separate sets of 2 million random samples

$$\begin{aligned} \text{From Theorem} \quad &: \quad .00059128 \times 2 \times 10^6 = 1182.56 \\ \text{From Random Samples} \quad &: \quad 1183,\ 1181,\ 1194,\ 1125,\ 1195 \end{aligned}$$

⇒ We can then compute the number of models with E_6 on v_{E_6} given this triangulation:

$$\text{Number of } E_6 \text{ Models on } T = .00059128 \times \frac{1}{3} \times 2.96 \times 10^{755} = 5.83 \times 10^{751}$$

In Lieu of a Summary – Two Invitations

Workshop on Machine Learning and the String Landscape
November 30-December 2, 2017

Northeastern University

In Lieu of a Summary – Two Invitations

String Pheno 2020

Northeastern University

THANK YOU!