Supervised learning of arithmetic invariants

Thomas Oliver
University of Nottingham

Symbolic Computation research theme
29 September 2021

Contents

Motivational context

Basic number field invariants

Polynomial	Degree	Rank	Description
$x^2 - 2$	2	1	Real quadratic
$x^2 + 2$	2	0	Imaginary quadratic
$x^4 - x^3 - 3x^2 + x + 1$	4	3	Totally real
$x^3 - x^3 - x^2 + x + 1$	4	1	Totally imaginary
$x^6 - 7x^4 + 14x^2 - 7$	6	5	Totally real
$x^6 - x^3 + 1$	6	2	Totally imaginary

Signature

The signature of F is (r_1, r_2), where r_1 (resp. r_2) is the number of real roots (resp. conjugate pairs of complex roots). We have $\mathrm{degree}(F/\mathbb{Q}) = r_1 + 2r_2$, and $\mathrm{rank}\left(\mathcal{O}_F^\times\right) = r_1 + r_2 - 1$.

Real quadratic fields, I

General form

For squarefree $d \in \mathbb{Z}_{>0}$, we have $\mathbb{Q}(\sqrt{d}) = \{a + b\sqrt{d} : a, b \in \mathbb{Q}\}$.

Failure of unique factorization

In $\mathbb{Q}(\sqrt{10})$, we have $9 = 3^2 = (7 - 2\sqrt{10})(7 + 2\sqrt{10})$.

Class number

The class number h_F of a number field F quantifies the failure of $\mathcal{O}_F$ to be a unique factorization domain.

Real quadratic fields, II

Class number	$\mathbb{Q}(\sqrt{d})$
1	$d = 2, 3, 5, 6, 7, 11, \ldots, 5581, \ldots, 2000029, \ldots$
2	$d = 10, 15, 65, 85, \ldots, 5133, \ldots$
3	$d = 79, 142, 229, 257, \ldots, 5081, \ldots$
4	$d = 82, 145, 445, 505, \ldots, 5545, \ldots$
$\vdots$	$\vdots$
21	$d = 7057, 13698, 49033, \ldots$
$\vdots$	$\vdots$

Class number problem (Gauss, 1801)

Does $\mathbb{Q}(\sqrt{d})$ have class number 1 for infinitely many squarefree d?

Elliptic curves

Theorem (Mordell, 1922)

The rational points on an elliptic curve defined over $\mathbb{Q}$ form a finitely generated abelian group.

Theorem (Mazur, 1977/78)

There are 15 explicit possibilities for the torsion subgroup of $E(\mathbb{Q})$.

Rank	Example	Proportion (conjectural)
0	$y^2 + y = x^3 - x^2$	50%
1	$y^2 + y = x^3 - x$	50%
2	$y^2 + y = x^3 + x^2 - 2x$	0%

Ranks of elliptic curves

Question (possibly dating back to Poincaré, 1901)

Which integers r can occur as the rank of an elliptic curve over $\mathbb{Q}$? Is the set of such r bounded?

Conjecture (Birch and Swinnerton-Dyer, 1960s)

The rank r is equal to the order of vanishing for the elliptic L-function $L(E, s)$ at $s = 1$. Furthermore, there is a formula for the leading Taylor coefficient involving certain invariants of E.

Supervised learning methodology

Experimental strategy

1. Given a set U of vectors and, for each $u \in U$, a label $\ell(u)$, consider the labeled dataset $\mathcal{D} = \{u \to \ell(u) : u \in U\}$.
2. Choose a subset $\mathcal{T} \subset \mathcal{D}$ and denote its complement by $\mathcal{V} = \mathcal{D} - \mathcal{T}$. We will refer to $\mathcal{T}$ as the training dataset, and $\mathcal{V}$ as the validation dataset.
3. Train a classifier on the set $\mathcal{T}$ with a standard supervised-learning algorithm.
4. For $u \in \mathcal{V}$, ask the classifier to determine $\ell(u)$. We record the precision and confidence.
5. Repeat steps 2 to 4 for different choices of $\mathcal{T}$. Record precision/confidence representative of several repetitions.

Naive implementation

Object	Vectors	Labels
number fields	defining polynomial	class number
elliptic curves	Weierstrass equation	rank

Experimental outcome

Results vary depending on specifics, but typically no better than guesswork.

Basic objective

Find better training vectors!

L-functions

L-functions

Dirichlet series $L(s) = \sum_{n=1}^{\infty} a_n n^{-s}$, converging for $\mathrm{Re}(s) \gg 0$. In this talk, it will be the case that $a_n \in \mathbb{Z}$.

Training vectors

Given $M \in \mathbb{Z}_{>0}$, and an L-function $L(s)$, we define the vector $L_M = (a_1, \ldots, a_M) \in \mathbb{Z}^M$.

Complexity

Measured by conductor $Q_L \in \mathbb{Z}$, which appears in a functional equation satisfied by L.

Experimental strategy, revisited

1. Choose a finite set $\mathcal{F}$ of L-functions and, for each $L \in \mathcal{F}$, let $I(L)$ denote an invariant of interest. Generate a labeled dataset of the form $\mathcal{D} = \{L_M \to I(L) : L \in \mathcal{F}\}$
2. Choose a subset $\mathcal{T} \subset \mathcal{D}$ and denote its complement by $\mathcal{V} = \mathcal{D} - \mathcal{T}$. We will refer to $\mathcal{T}$ as the training dataset, and $\mathcal{V}$ as the validation dataset.
3. Train a classifier on the set $\mathcal{T}$ with a standard supervised-learning algorithm.
4. For $L \in \mathcal{V}$, ask the classifier to determine $I(L)$. We record the precision and confidence.
5. Repeat steps 2 to 4 for different choices of $\mathcal{T}$. Record precision/confidence representative of several repetitions.

Typical specifications

Vector length

$M \approx 10^2$.

Labeled dataset

$\mathcal{D} = \{L_M \to I(L) : B_0 < Q_L < B_1\}$, with conductor bounds B_0, B_1 chosen so that $10^3 < |\mathcal{D}| < 10^6$.

Training/Validation ratio

$$\frac{|\mathcal{T}|}{|\mathcal{D}|} \in \{0.2, 0.8, 0.7\}.$$

Case study 1: Number fields

Dedekind zeta functions

Dedekind zeta function

$$\zeta_F(s) = \prod_{\mathfrak{p}} \left(1 - N(\mathfrak{p})^{-s}\right)^{-1} = \sum_{I \leq \mathcal{O}_F} N(I)^{-s},$$

where $\mathfrak{p}$ varies over prime ideals in $\mathcal{O}_F$, I varies over the non-zero ideals in $\mathcal{O}_F$.

Dirichlet coefficients

We have $\zeta_F(s) = \sum_{n=1}^{\infty} a_n n^{-s}$, where

$$a_n = \#\{N(I) = n : I \leq \mathcal{O}_F\}.$$

The conductor Q_F is equal to the discriminant, which is divisible exactly by the primes ramified in F.

Promising example: class numbers

- Let $\mathcal{F} = \{[F : \mathbb{Q}] = 2, h_F \in \{1, 2\}, 1 < \Delta_F < 10^6\}$.
- Labeled dataset $\mathcal{D}_Z = \{(a_1, \ldots, a_M) \to h_F : F \in \mathcal{F}\}$.
- Randomly splitting $\mathcal{D}_Z$ into a disjoint union $\mathcal{T} \coprod \mathcal{V}$, a random forest classifier trained on $\mathcal{T}$ predicts the (unseen) class numbers of $\mathcal{V}$ with accuracy ≈ 0.96 and confidence ≈ 0.92.
- Furthermore, we find that the same trained classifier can predict class numbers for larger discriminants with decent levels of success.

$[T_0, T_1]$	$[V_0, V_1]$	Precision	Confidence
$[1, 1 \times 10^6]$	$[1, 1 \times 10^6]$	0.96	0.92
$[1, 1 \times 10^6]$	$[1 \times 10^6, 2 \times 10^6]$	0.92	0.86
$[1, 1 \times 10^6]$	$[2 \times 10^6, 3 \times 10^6]$	0.91	0.84

Table: Training discriminants in range $[T_0, T_1]$, validation discriminants in range $[V_0, V_1]$.

Cautionary tale: Signatures

- Let $\mathcal{F}$ consist of cyclic extensions with bounded discriminant, degree 6 and unit group rank r_F equal to either 5 (totally real) or 2 (totally imaginary).
- Applying standard classifiers to the labeled dataset $\{(a_1, \ldots, a_M) \to r_F : F \in \mathcal{F}\}$ leads to predictions with accuracy < 0.6.
- By way of contrast, let $\mathcal{F}$ contain cyclic extensions of degree $d_F \in \{4, 6, 8\}$. Training on a labelled dataset $\{(a_1, \ldots, a_M) \to d_F : F \in \mathcal{F}\}$, a classifier is able to make predictions with accuracy ≈ 0.98 and precision ≈ 0.97.
- Recall that the extension degree is equal to $r_1 + 2r_2$, and the unit group rank is equal to $r_1 + r_2 - 1$.

Naive training for signatures

Instead of zeta coefficients, one can use defining polynomials:

$$P(x) = x^n + c_{n-1}x^{n-1} + \cdots + c_1 x + c_0, \quad c_i \in \mathbb{Z}, \quad n = [F : \mathbb{Q}].$$

We introduce the following labeled dataset:

$$\mathcal{D}_P = \{(c_0, \ldots, c_{n-1}) \to r_F : F \in \mathcal{F}\}.$$

Trained on this data, a random forest classifier can make much more accurate rank predictions.

Galois group	(r_1, r_2)	rank$(\mathcal{O}_F^\times)$	$\mathcal{D}_P$ precision	$\mathcal{D}_P$ confidence
C_6	(6,0) (0,3)	5 2	0.97	0.93
C_8	(8,0) (0,4)	7 3	> 0.99	> 0.99
D_4	(8,0) (0,4)	7 3	0.98	0.95

Logistic regression example, I

Rounding function

Given a real number X, let $[X]$ denote the integer nearest to X.

Logistic sigmoid

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

Objective

Let $\mathcal{F}$ contain cyclic degree 6 number fields with rank 2 or rank 5 and bounded conductor. We aim to find $(w_0, \ldots, w_5) \in \mathbb{R}^6$ such that the function $3[\sigma(c_0 w_0 + \cdots + c_4 w_4 + c_5 w_5)] + 2 \in \{2, 5\}$ predicts the rank.

Logistic regression example, II

Best fit

We get 94% precision with best fit:

$$-0.000169037c_0 - 0.0000689721c_1 - 0.000120625c_2 \\ - 0.00196535c_3 - 0.058735c_4 + 0.917924c_5.$$

The accuracy of this model varies with r_F and Δ_F. More precisely, the model predicts rank 2 with accuracy > 0.91 for almost all ranges of $|\Delta_F|$ occurring in the dataset with overall precision 0.99. On the other hand, the model above performs poorly for rank 5 fields with $|\Delta_F| < 1.80 \times 10^9$ (around 77% accuracy) and $|\Delta_F| > 2.55 \times 10^{14}$ (around 60% accuracy); the overall precision for the classification of rank 5 fields is 0.89.

Case study 2: Hyperelliptic curves

Hasse–Weil L-functions

Local zeta function

Let X be a smooth, projective, geometrically connected curve of genus g. For each good prime p of X, we define:

$$Z(X/\mathbb{F}_p; T) = \exp\left(\sum_{k=1}^{\infty} \frac{\#X\left(\mathbb{F}_{p^k}\right) T^k}{k}\right).$$

Theorem (Weil, 1949)

$$Z(X/\mathbb{F}_p; T) = \frac{L_p(X, T)}{(1-T)(1-pT)},$$

where $L_p(T) \in \mathbb{Z}[T]$ has degree $2g$ and constant term 1.

Elliptic L-functions

Elliptic Euler factors

If E is an elliptic curve defined over $\mathbb{Q}$ and p is a good prime, then $L_p(E,T) = 1 - a_p T + pT^2$, where $a_p = p + 1 - \#E(\mathbb{F}_p)$. For a bad prime p, we also define a_p this way.

Training vectors

For $i \in \mathbb{Z}_{>0}$, let p_i denote the ith prime. For $M \in \mathbb{Z}_{>0}$, we introduce the vector:

$$v_L(E) = (a_{p_1}, \ldots, a_{p_M}) \in \mathbb{Z}^M.$$

Conductor Q_E divisible only by bad primes.

Elliptic curve ranks

Q_E training	M	Q_E validation	P	C
$[1, 1 \times 10^4]$	100	$[1, 1 \times 10^4]$	0.98	0.96
"	300	"	0.99	0.98
$[2 \times 10^4 + 1, 3 \times 10^4]$	300	$[2 \times 10^4 + 1, 3 \times 10^4]$	0.96	0.92
"	500	"	0.97	0.94
$[1, 1 \times 10^4]$	300	$[2 \times 10^4 + 1, 3 \times 10^4]$	0.92	0.85

Table: The precision and confidence of a logistic regression classifier when asked to distinguish elliptic curves over $\mathbb{Q}$ with rank 0 from those with rank 1. The classifier is trained on the conductor range specified by the first column, using the number of Euler factors given in the second column, and verified on the conductor range indicated by the third column.

Genus 2 L-functions

Euler factors

If $X = C$ is a smooth projective geometrically connected genus 2 curve defined over $\mathbb{Q}$ and p is a good prime for C, then:

$$L_p(C, T) = 1 + a_{1,p}T + a_{2,p}T^2 + a_{1,p}pT^3 + p^2T^4.$$

For a bad prime p, we put $(a_{1,p}, a_{2,p}) = (0, p)$.

Training vectors

For a positive integer M, we introduce the vector:

$$v_L(C) = \left((a_{1,p_2}, a_{2,p_2}), \ldots, (a_{1,p_{M+1}}, a_{2,p_{M+1}})\right) \in \left(\mathbb{Z}^2\right)^M.$$

Genus 2 ranks

Theorem (Weil, 1929)

The Jacobian of a genus 2 curve is a 2-dimensional abelian variety, whose rational points form a finitely generated abelian group.

New phenomenon

A little under 1/3 of the (Jacobians of) genus 2 curves over $\mathbb{Q}$ on the LMFDB have rank 2.

Supervised learning experiment

A logistic regression classifier trained on $v_L(C)$ is able to distinguish between curves of ranks 0, 1, and 2 for conductors in the range $1 < Q_C < 10^6$, with precision ≈ 0.97 and confidence ≈ 0.96.

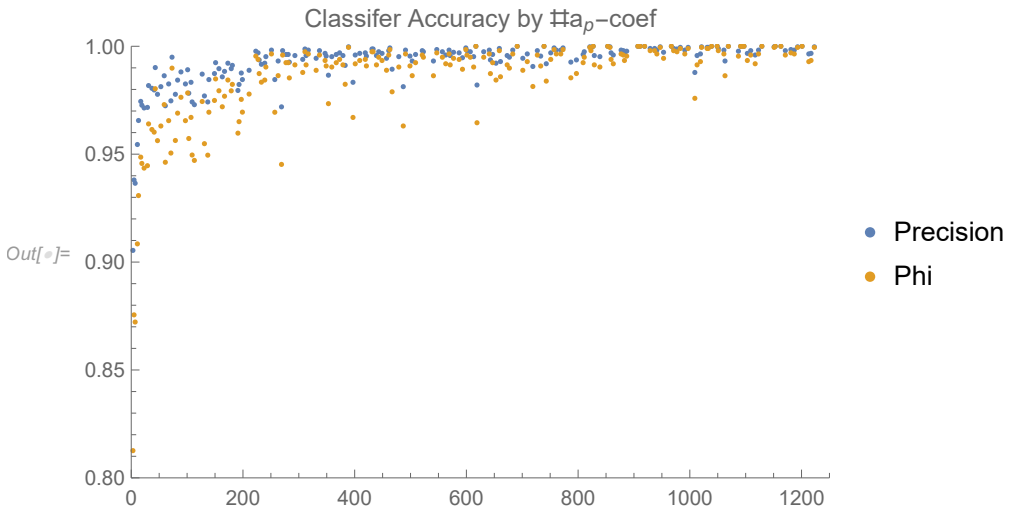
Case study 3: Random matrix training

Endomorphisms of elliptic curves

- A generic elliptic curve over $\mathbb{Q}$ has endomorphism ring isomorphic to $\mathbb{Z}$.
- The Sato–Tate conjecture (proved over totally real fields) implies that the Euler factors are distributed the same way for all generic curves. More precisely, they are distributed like the characteristic polynomials of random matrices in $\mathrm{SU}(2)$.
- The non-generic elliptic curves are those with CM. In this case, the Euler factors are associated with Hecke characters and distributed differently. This time the distribution matches random matrices in the normalizer of $\mathrm{U}(1)$ in $\mathrm{SU}(2)$.

CM elliptic curves

A naive Bayes classifier trained on random matrices and validated on elliptic curves is able to distinguish between CM and non-CM curves. In the following image the precision and confidence is plotted against the number of random matrices used in training.

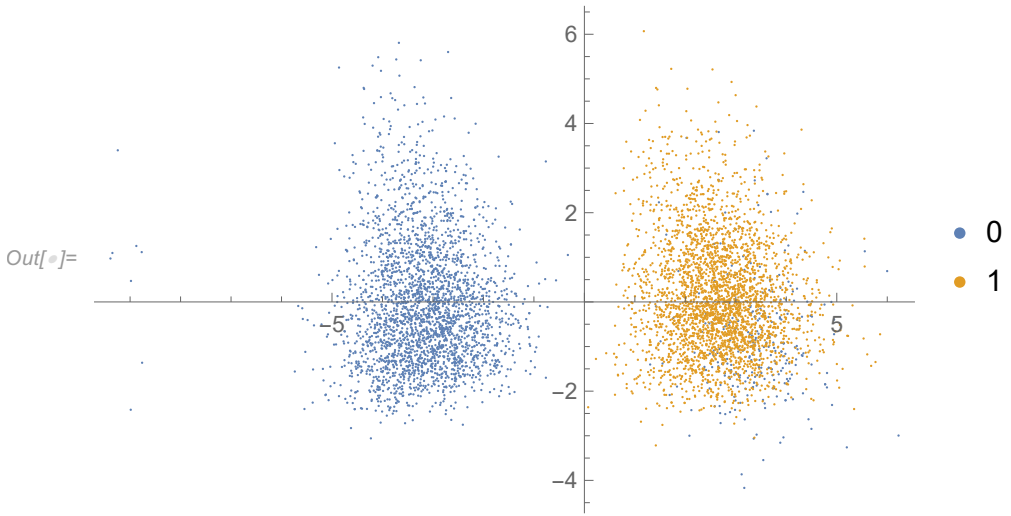


Generic genus 2 curves: Supervised

- According to the generalized Sato–Tate conjecture, Euler factors of generic genus 2 curves are distributed like random matrices in $\mathrm{USp}(4)$.
- Again a naive Bayes classifier trained on random matrices can separate the generic and non-generic case.
- There are 54 non-generic cases (33 over $\mathbb{Q}$), which are subgroups of $\mathrm{USp}(4)$ satisfying certain axioms. We will touch upon the non-generic case again below.

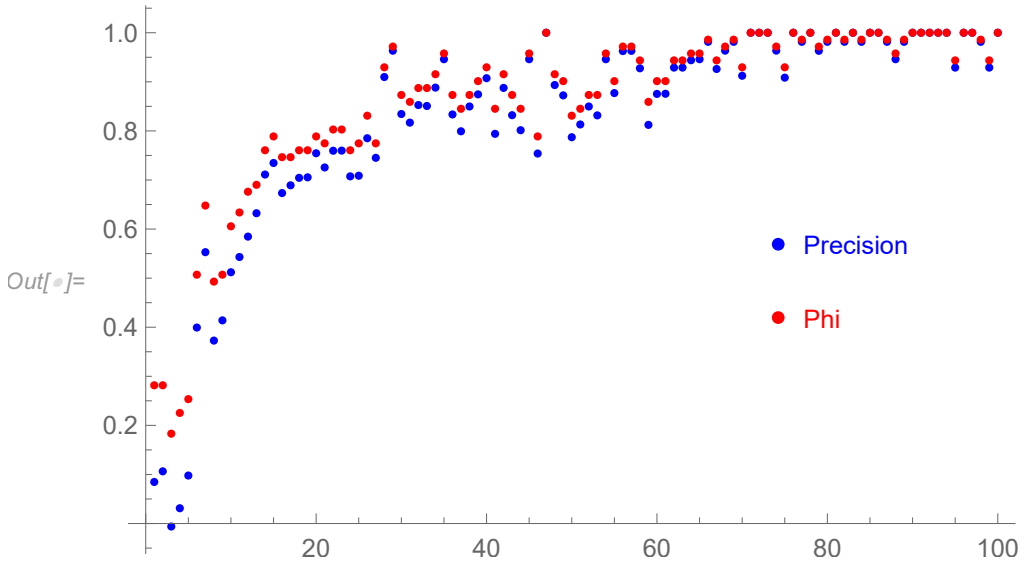
Generic genus 2 curves: Unsupervised

A two-dimensional projection of labeled coefficient pairs in $\mathbb{R}^{400}$ corresponding to generic and non-generic curves gives the following image:



Non-generic genus 2 curves

There are 33 possible non-generic Sato–Tate groups for genus 2 curves over $\mathbb{Q}$. The image belows shows the performance of a naive Bayes classifier in distinguishing groups $J(E_n)$, $n \in \{1, 2, 3, 4, 6\}$, when trained on random matrices and validated on curves, relative to the number of Euler factors used.



Adverts

Adverts

Preprints

- He–Lee–O, *Machine-learning the Sato–Tate conjecture*, arXiv:2011.08958.
- He–Lee–O, *Machine-learning number fields*, arXiv:2011.08958.
- He–Lee–O, *Machine-learning arithmetic curves*, arXiv:2012.04084.

More ML in pure maths

DANGER (Data, numbers and geometry), `https://sites.google.com/view/danger-workshop`.