Quantitative machine learning study of the critical 2D Ising model

Davide Vadacchino[1]

(with B. Lucini [2] and C. Giannetti [3])

[1] INFN - Sezione di Pisa

[2] Mathematics Department - Swansea University - UK

[3] College of Engineering - Swansea University - UK

XVII workshop on Statistical Mechanics and non Perturbative Field Theory,
14 December 2017

Introduction
Why Machine Learning?

- Large Variety of uses: spam filters, personalized ads, shopping assistance, face recognition, Health Sciences, ... [Domany, session 1]
- In general: pattern recognition, classification.
- Algorithms: (deep) neural networks, support vector machines, . . .
- Many ready to use libraries in a variety of programming languages: scikit-learn, tensorFlow, Theano, ... [Chang, Chih-Chung and Lin, Chih-Jen, 2011]
- Several studies of ML applied to the study of phase transition are already present in the litterature. [Melko, Rogers, Carrasquilla and many others]

Introduction
Summary

Our question...

Can we obtain the critical indices and critical temperature of the 2D Ising model using a Support Vector Machine?

- We make only one assumption: that there is a (second order) phase transition somewhere in T.
- We choose to study the 2D Ising model because it is exactly solved and critical slowing down can be overcome with cluster algorithms. see for example [Wolff, '89]
- We want to use one of the simplest and most transparent example of supervised learning algorithm: a Support Vector Machine. [V. N. Vapnik, A. Y. Chervonenkis '63]
- We perform the standard multihistogram analysis on data obtained from simulation to compare with our results. [Ferrenberg, Swenden '88]

Linear SVM

Statement of the problem

Given a set of *training* data

$$(\vec{x}_1, y_1), (\vec{x}_2, y_2), \cdots, (\vec{x}_N, y_N) \tag{1}$$

where $\vec{x}_i \in \mathrm{R}^p$ and $y_i = \pm 1$ labels the class. We want our machine f to classify any additional data set we feed it

$$f(\vec{x}_{N+1}) = y_{N+1} \tag{2}$$

Support Vector Machine

The SVM method seeks to find the maximum margin hyperplane defined by

$$\vec{\omega} \cdot \vec{x} - b = 0 \tag{3}$$

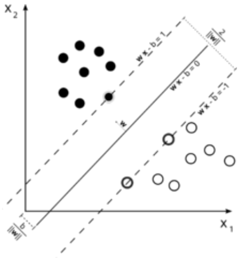
that has the largest possible distance from either of the two classes:

- $\vec{\omega}$ is the normal to the plane in $\mathbb{R}^p$.
- $b/||\vec{\omega}||$ is the offset with respect to the origin.

Linear Classification

If the samples are linearly classificable, they are separated by a *margin*, bounded by the planes defined by

$$\vec{\omega} \cdot \vec{x} - b = -1, \qquad \vec{\omega} \cdot \vec{x} - b = 1 \tag{4}$$



- The margin has size $2/||\vec{\omega}||$.
- On either side of the margin,

$$y_i \left(\vec{\omega} \cdot \vec{x} - b \right) \geq 1.$$

- Samples on the margin define the support vectors.
- For a sample that falls into the margin

$$y_i \left(\vec{\omega} \cdot \vec{x} - b \right) \leq 1.$$

Solution

The problem is solved once $\vec{\omega}$ and b are found. Then

$$f(\vec{x}) = \text{sign}\left(d\left(\vec{x}\right)\right)$$

provides the classification, where $d(\vec{x}) = \vec{\omega} \cdot \vec{x} - b$ is the decision function.

SVM as a minimization problem

Primal problem

Minimize L,

$$L = \frac{1}{2}||\vec{\omega}||^2 + C\sum_{i=1}^{N}\zeta_i + \sum_{i=1}^{N}\alpha_i\left(1 - y_i\left(\vec{\omega}\cdot\vec{x}_i - b\right)\right) - \sum_{i=1}^{N}\gamma_i\zeta_i \tag{5}$$

where α_i, γ_i are Lagrange multipliers, and C is a regularization parameter.

Dual problem

Minimizing L w.r.t $\vec{\omega}$, b and ζ_i yelds the *quadratic program*

$$\min_{\alpha}\frac{1}{2}\alpha^T H\alpha - \alpha^T e$$
$$\text{s. t. } \alpha^T y = 0, \qquad 0 \leq \alpha_i \leq C$$

where now $\vec{\omega} = \sum_i y_i\alpha_i\vec{x}_i$ and

$$H_{ij} = y_i y_j\ \vec{x}_i\cdot\vec{x}_j \tag{6}$$

Nonlinear classification

Feature mapping

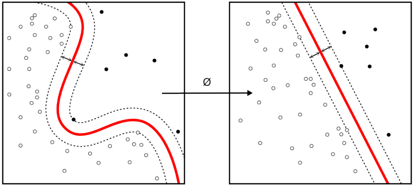
In some cases, problems that don't accept a linear classification in $\vec{x}_i$ might accept one in $\phi(\vec{x}_i)$ in some other space. The calculations are the same and lead to

$$\min_{\alpha} \frac{1}{2} \alpha^T H \alpha - \alpha^T e$$
$$\text{s. t. } \alpha^T y = 0, \qquad 0 \leq \alpha_i \leq C$$

where now $\vec{\omega} = \sum_i y_i \alpha_i \; \phi(\vec{x}_i)$ and

$$H_{ij} = y_i y_j \; \mathcal{K} \left(x_i , \; x_j \right) \tag{7}$$

where $\mathcal{K} = \phi(\vec{x}_i) \cdot \phi(\vec{x}_j)$ is the kernel.



Nonlinear classification

To help intuition...

- A polynomial kernel of degree d, $\mathcal{K} = (c_o + \vec{x}_i \cdot \vec{x}_j)^d$ produces as features the d point correlation functions of the system. For example, for $d = 2$

$$\phi(\vec{x}) = (x_1^2, \cdots, x_{L^2}^2, \sqrt{2}\, x_0 x_1, \cdots, \sqrt{2}\, x_{L^2-1} x_{L^2}) \tag{8}$$

- The decision function is now

$$d(\vec{x}) = \sum_{i=1}^{N} y_i \alpha_i \mathcal{K}\left(\vec{x}_i, \vec{x}\right) - b \tag{9}$$

  Its sign determines the classification, its value is the distance of $\phi(\vec{x})$ from the maximum margin hyperplane.

- The accuracy of classification can be computed for a sample for which the labeling is known,

$$\mathrm{acc}_k = \frac{\#\text{ correctly classified with label k}}{\#\text{ total data in sample}} \tag{10}$$

The 2D Ising model
Definition and Numerical setup

The (ferromagnetic) Ising model is defined by the Hamiltonian,

$$\mathcal{H} = -J \sum_{\langle i,j \rangle} \sigma_i \sigma_j, \qquad J > 0 \tag{11}$$

where $\langle i,j \rangle$ denotes the sum over next neighbours and $\sigma_i = \pm$.
For this study, a square lattice and $D = 2$, then the model is exactly solved and :

- The order parameter associated to the transition is

$$m = \frac{1}{L^2} \sum_{i=1}^{N} \sigma_i \tag{12}$$

- At $T_c = 2/\ln\left(1+\sqrt{2}\right)$, there is a second order phase transition with exponents $\nu = 1$ and $\gamma = 7/4$. Using the hyperscaling relations, all the other exponents can be computed.
- $N = 200$ decorrelated configurations were generated using the Wolff cluster algorithm on $L \times L$ lattices, with $L = 128, 240, 360, 440, 512, 760, 1024$.

The 2D Ising model
Training the SVM

We want the SVM to classify raw configurations as being ordered or disordered:

- We place the original spins at temperature T_k in a L^2 components vector,

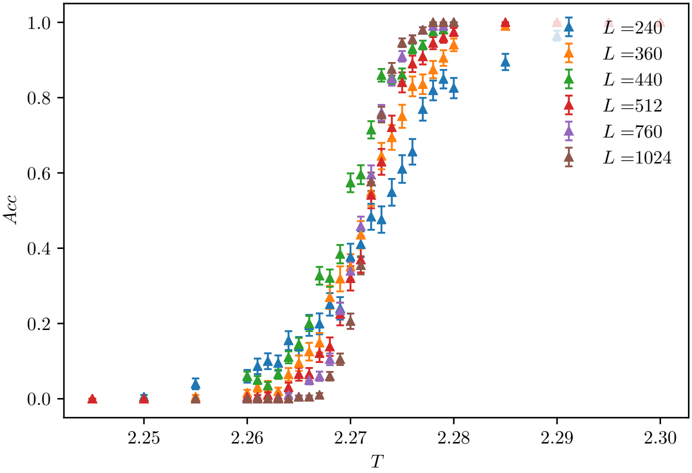
$$\vec{x}_i^k = (\sigma_0, \sigma_1, \cdots, \sigma_{L^2})_i^k \tag{13}$$

  where i labels the configuration, $i = 1, \cdots, 200$, k the temperatures.
- We associate the labels -1 and $+1$, respectively, to the ordered and disordered phase.
- Bayesian inference techniques suggest that the quadratic kernel is the optimal choice across the set of the most popular ones (polynomial and gaussian).
- Let T_o and T_d be, respectively, the ordered and disordered training temperatures. We adopt a self consistent procedure to obtain T_o and T_d:

$$T_o = \max_k \left(T_k \quad / \quad \mathrm{acc}_o\left(\left\{\vec{x}_i^k\right\}\right) = 0\right)$$
$$T_d = \min_k \left(T_k \quad / \quad \mathrm{acc}_o\left(\left\{\vec{x}_i^k\right\}\right) = 1\right)$$

  and we visualize the accuracy to classify the configurations as disordered at all the other temperatures.

The 2D Ising model
Classification scores



The 2D Ising model

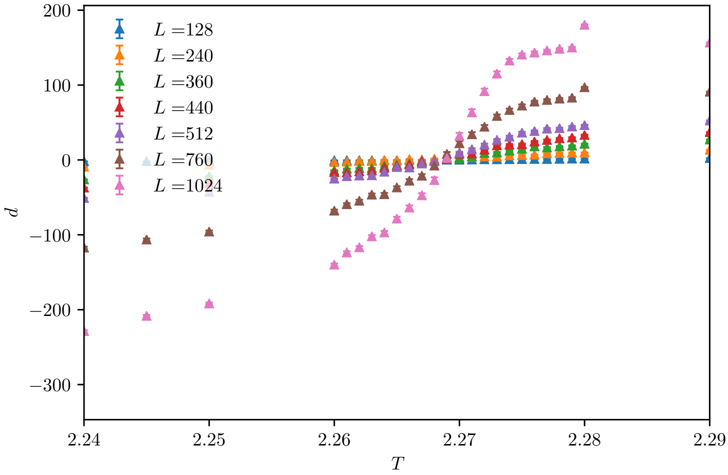
Closing on the critical behaviour

- The classification *sharpens* when L is increased.
- As shown in [Melko, Ponte 2017], for small C this selects a linear function of m^2 as a decision function.
- At each temperature T_k, we compute the average decision function and its error

$$\langle d \rangle = \frac{L^2}{200} \sum_{j=1}^{200} d(\vec{x}_j), \quad \chi_d = L^2 \sqrt{\frac{1}{200} \sum_{j=1}^{200} (d(\vec{x}_j) - \langle d \rangle)^2} \tag{14}$$

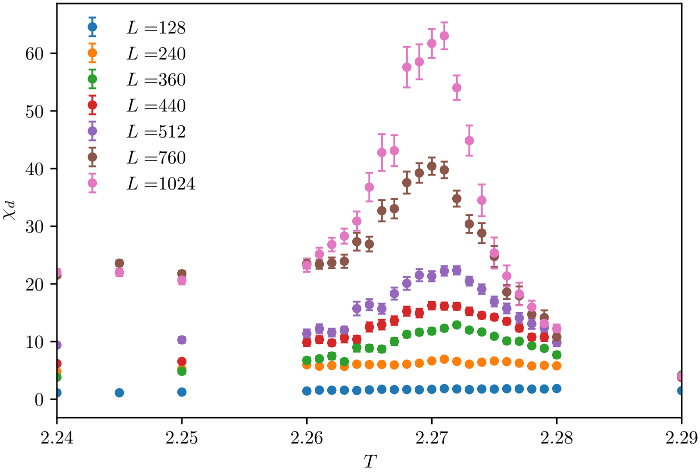
The 2D Ising model

Values of the decision function versus T



The 2D Ising model

Values of the error of decision function versus T



The 2D Ising model

Finite size scaling

What we observe

It seems that, at $T = T_c(L)$:

- $d \sim 0$.
- χ_d reaches its maximum value.

What we think we know [Preliminary]

Since d depends on m^2, we expect $\chi_{\mathrm{d,max}}(L)$ to scale as

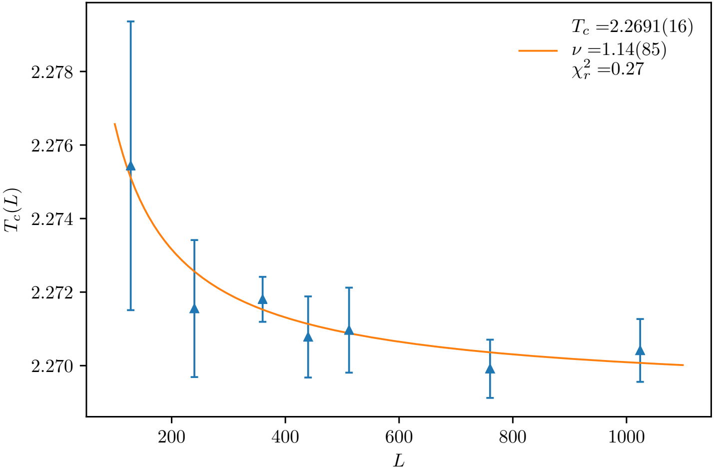
$$\chi_{\mathrm{d,max}}(L) \propto L^{2+\frac{\gamma/2-\beta}{\nu}} \quad (15)$$

while, for $T_c(L)$

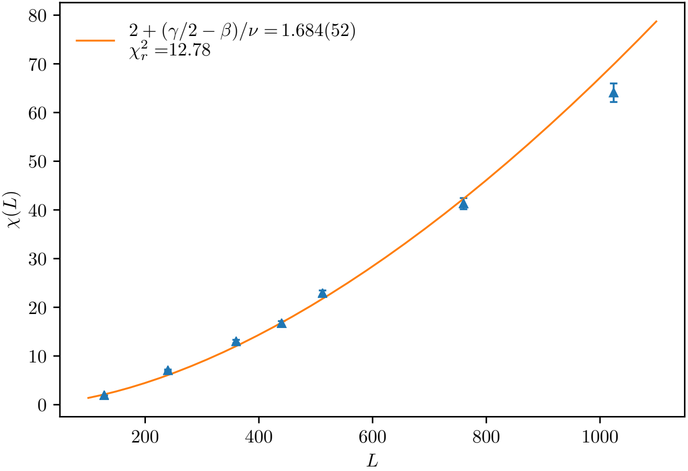
$$T_c(L) - T_c(\infty) \propto L^{-1/\nu} \quad (16)$$

We extract $T_c(L)$ and $\chi_{\mathrm{d,\,max}}(L)$ and fit the above scaling behaviour. We expect $T_c = 2.2692$, $\nu = 1$, $2 + (\gamma/2 - \beta)/\nu = 2.75$

Computation of T_c



Computation of $2 + (\gamma - \beta)/\nu$



Systematics

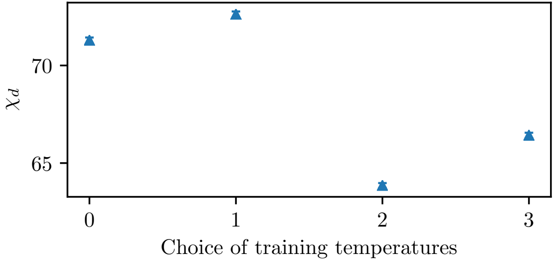
Where do the errors come from?

- From the $\vec{x}_i$: statistical, depends on how to configurations scatter.
- from the α_i's: systematic, depends on the choice of training temperatures. Heuristically...

$$\delta\, d = \frac{\delta d}{\delta \vec{x}}\; \delta\, \vec{x} + \sum_i \frac{\delta d}{\delta \alpha_i}\; \delta\, \alpha_i \tag{17}$$

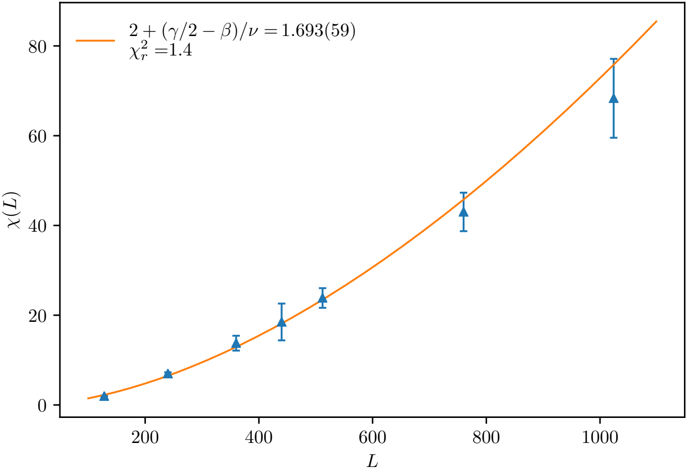
where α_i are determined during training, i.e. they depend on T_o and T_d.

- Arbitrary rescaling performed by libsvm in the scikit-learn package...



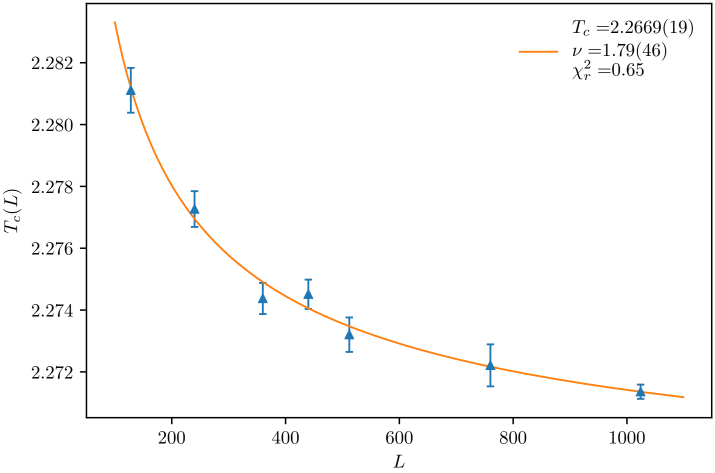
@L = 1024, for choices of T_o and T_d around the autoconsistent values.

Tentative results corrected for systematics



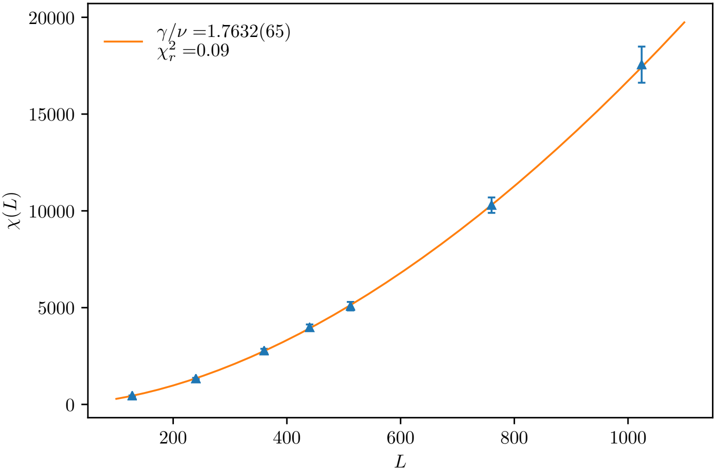
Multihistogram method

Determination of T_c and ν



Multihistogram method

Determination of γ/ν



PRELIMINARY

Method	T_c	ν	γ/ν	$2 + (\gamma/2 - \beta)/\nu$
Exact	$2.269619\ldots$	1.0	$7/4 = 1.75$	2.75
MH	2.2669(19)	1.79(46)	1.7632(65)	-
SVM	2.2691(16)	1.14(85)	-	1.693(59)

How is the difference between 2.75 and 1.693(59) explained?

- Rescaling performed in scikit-learn?
- Spurious scalings introduced in the training procedure?
- ...

Conclusion - PRELIMINARY

Conclusions

- T_c and ν can be estimated from finite size scaling.
- The difference between the naively predicted value of $2 + (\gamma/2 - \beta)/\nu$ and its measured value is ~ 1.
- The accuracy of these estimates is slightly worse than that obtained with standard techniques.

Future directions & Improvements

- Improve the estimation of the systematical error (especially the effects coming from the choice of the training temperatures...)
- Try on a model with a transition for which local order parameter cannot be identified.

Thank you for your attention