Machine learning applied to the one- and two dimensional Ising model.

FYS-STK4155: Project 2

Morten Ledum & Håkon Kristiansen
github.com/mortele/FYS-STK4155

November 1, 2018

Abstract

Contents

INTRODUCTION

There are many problems that require a probability estimate as output. This could for example predicting whether a person will develop a specific disease given genetic information. Another example, which we will examine closer, is to predict if a given spin-configuration generated from the two-dimension Ising model is ordrered or disordered. Problems of this type are referred to as *classification* problems.

Classification is fundamentally different from the regression problems we studied previously, in the sense that the predicted outcome only takes values across discrete categories. Thus, we will need different tools than that of linear regression. In this work we first consider *logistic regression* as a method for classification.

Somethingsomething background deep neural networks

Apply it to one-dim Ising and classification of two-dim Ising.

THEORY

In the following we outline the theory of the present work. We consider logistic regression as a model for classification problems. Furthermore, neural networks are discussed both in the context of regression analysis and classification.

The theoretical aspects of linear regression have been discussed in previous work and is not repeated here.

In contrast to the linear regression model, we can not find the optimal parameters of the logistic or neural network models analytically. Thus, we have ro rely on numerical methods for optimization. In particular we will give a brief summary gradient descent methods.

Logistic regression

Suppose that we are given a dataset $\{(\mathbf{x}^{(i)}, y_i)\}_{i=1}^n$ where we have p predictors for each data sample $\mathbf{x}^{(i)} = \{x_1^{(i)}, \cdots, x_p^{(i)}\}$. The responses/outcomes y_i are discrete and can only take values from $k = 0, 1, \ldots, K-1$ (i.e K classes). The goal is to predict the output classes given n samples each containing p predictors. Throughout this section we assume that there are just two possible outcomes, i.e $y_i \in \{0, 1\}$.

In logistic regression, in contrast to linear regressions, we model the *probability* that y_i belongs to class 1, given $\mathbf{x}^{(i)}$. Let $p(y|x)$ denote the probability of event y given x, then the *logistic model* is

$$p\left(y = 1|\mathbf{x}; \beta\right) = \frac{1}{1 + e^{-\beta \cdot \mathbf{x}}} \tag{1}$$

$$p\left(y = 0|\mathbf{x}; \beta\right) = 1 - p\left(y = 1|\mathbf{x}; \beta\right). \tag{2}$$

Here $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ are the parameters of the model. Note the appearance of the intercept term β_0. In order to keep notation compact $\mathbf{x}^{(i)}$ can be augmented to incorporate the intercept by adding a 1 to each sample, i.e

$$\mathbf{x}^{(i)} \to \{1, x_1^{(i)}, \ldots, x_p^{(i)}\}.$$

The term $\beta \cdot \mathbf{x} = \beta_0 + \sum_{k=1}^p \beta_k x_k$ is known as the *log-odds* and the function

$$\sigma(\beta \cdot \mathbf{x}) = \frac{1}{1 + e^{-\beta \cdot \mathbf{x}}} \tag{3}$$

is called the *sigmoid* of $\beta \cdot \mathbf{x}$. Also note that the sigmoid satisfies

$$\lim_{t \to \infty} \sigma(t) = 1 \tag{4}$$

$$\lim_{t \to -\infty} \sigma(t) = 0 \tag{5}$$

which justifies its use as a model for probabilities.

The logistic model can now be used for classification by predicting a class using the estimated probabilities according to

$$\hat{y}_i = \begin{cases} 1 & \text{if } p\left(y = 1|\mathbf{x}^{(i)}\right) \geq 0.5 \\ 0 & \text{if } p\left(y = 1|\mathbf{x}^{(i)}\right) < 0.5. \end{cases} \tag{6}$$

Training the logistic model

How do we train the logistic model? The answer is to use the prinicple of *maximum likelihood*. Under the assumption that every sample $\mathbf{x}^{(i)}$ is indenpendent, the likelihood is given by

$$\begin{aligned} L(\beta) &= \prod_{i:y_i=1} p(y_i = 1|\mathbf{x}^{(i)}) \prod_{i:y_i=0} p(y_i = 0|\mathbf{x}^{(i)}) \\ &= \prod_{i=1}^n p(y_i = 1|\mathbf{x}^{(i)})^{y_i} (1 - p(y_i = 1|\mathbf{x}^{(i)}))^{1-y_i}. \end{aligned} \tag{7}$$

Then, the parameters β are chosen to maximize the likelihood.

It turns out that it is easier to work with the *log-likelihood*

$$\begin{aligned} l(\beta) &= \log(L(\beta)) \\ &= \sum_{i=1}^n y_i p(y_i = 1|\mathbf{x}^{(i)}) + (1 - y_i)(1 - p(y_i = 1|\mathbf{x}^{(i)}). \end{aligned} \tag{8}$$

Maximizing the logarithm of a function is equivalent to maximizing the function itself.

In order to see this, let $f(x)$ be a real valued function and let x^* be a maximum point of $f(x)$, i.e

$$f'(x^*) = 0, \quad f''(x^*) < 0. \tag{9}$$

Furthermore, assume that $f(x) > 0$ and consider $\log(f(x))$. Taking derivatives we have that

$$\frac{d}{dx} \log(f(x)) = \frac{f'(x)}{f(x)} \tag{10}$$

$$\Rightarrow \frac{d}{dx} \log(f(x^*)) = 0 \tag{11}$$

$$\frac{d^2}{dx^2} \log(f(x)) = \frac{f''(x)f(x) - f'(x)^2}{f(x)^2} \tag{12}$$

$$\Rightarrow \frac{d^2}{dx^2} \log(f(x^*)) < 0, \tag{13}$$

where the last inequality follows from the fact that we assumed $f'(x^*) = 0$, $f''(x^*) < 0$ and $f(x) > 0$. Hence, x^* also maximize $\log(f(x))$.

Thus, taking β to maximize the log-likelihood is equivalent to maximizing the likelihood itself.

Finally, we take our cost function to be the so-called *cross-entropy* which is defined as the negative log-likelihood

$$C(\beta) \equiv= -l(\beta). \tag{14}$$

Then, β is found by *minimizing* the cross-entropy.

Note here that we can not find a analytical solution for the maximizer. This means that we have to use a numerical opimization algorithm, such as gradient descent which we discuss later, to find the optimal parameters.

Neural networks

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Gradient Descent

Almost every problem in machine learning and data science starts with a dataset X, a model $g(\theta)$, which is a function of the parameters θ and a cost function $C(X, g(\theta))$ that allows us to judge how well the model $g(\theta)$ explains the observations X. The model is fit by finding the values of θ that minimize the cost function. Ideally we would be able to solve for θ analytically, however this is not possible in general and we must use numerical methods to compute the minimum.

The method of steepest descent

The basic idea of gradient descent is that a function $F(\mathbf{x})$, $\mathbf{x} \equiv (x_1, \cdots, x_n)$, decreases fastest if one goes from $\mathbf{x}$ in the direction of the negative gradient $-\nabla F(\mathbf{x})$. It can be shown that if

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \gamma_k \nabla F(\mathbf{x}_k), \quad \gamma_k > 0 \tag{15}$$

for γ_k small enough, then $F(\mathbf{x}_{k+1}) \leq F(\mathbf{x}_k)$. This means that for a sufficiently small γ_k we are always moving towards smaller function values, i.e a minimum.

This observation is the basis of the method of steepest descent, which is also referred to as just gradient descent (GD). One starts with an initial guess $\mathbf{x}_0$ for a minimum of F and compute new approximations according to

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \gamma_k \nabla F(\mathbf{x}_k), \quad k \geq 0. \tag{16}$$

The parameter γ_k is often referred to as the step length or the learning rate in the context of ML.

Ideally the sequence $\{\mathbf{x}_k\}_{k=0}$ converges to a *global* minimum of the function F. In general we do not know if we are in a global or local minimum. In the special case when F is a *convex function, all local minima are also global minima*, so in this case gradient descent can converge to the global solution. The advantage of this scheme is that it is conceptually simple and straightforward to implement.

However the method in this form has some severe limitations:

- In machine learing we are often faced with non-convex high dimensional cost functions with many local minimum. Since GD is deterministic we will get stuck in a local minimum, if the method converges, unless we have a very good intial guess. This also implies that the scheme is sensitive to the chosen initial condition.
- Note that gradient is a function of $\mathbf{x} = (x_1, \cdots, x_n)$ which makes it expensive to compute numerically.
- GD is sensitive to the choice of learning rate γ_k. This is due to the fact that we are only guaranteed that $F(\mathbf{x}_{k+1}) \leq F(\mathbf{x}_k)$ for *sufficiently* small γ_k. The problem is to determine an optimal learning rate. If the learning rate is chosen to small the method will take a long to converge and if it is to large we can experience erratic behavior.
- Many of these shortcomings can be alleviated by introducing randomness. One such method is that of Stochastic Gradient Descent (SGD).

Stochastic Gradient Descent

Stochastic gradient descent (SGD) and variants thereof address some of the shortcomings of the Gradient descent method discussed above.

The underlying idea of SGD comes from the observation that the cost function, which we

want to minimize, can almost always be written as a sum over n datapoints $\{\mathbf{x}_i\}_{i=1}^n$,

$$C(\theta) = \sum_{i=1}^{n} c_i(\mathbf{x}_i, \theta). \tag{17}$$

This in turn means that the gradient can be computed as a sum over i-gradients

$$\nabla_\theta C(\theta) = \sum_{i}^{n} \nabla_\theta c_i(\mathbf{x}_i, \theta). \tag{18}$$

Now, stochasticity/randomness is introduced by only taking the gradient on a subset of the data called minibatches. If there are n datapoints and the size of each minibatch is M, there will be n/M minibatches. We denote these minibatches by B_k where $k = 1, \cdots, n/M$.

As an example, suppose we have 10 datapoints $(\mathbf{x}_1, \cdots, \mathbf{x}_{10})$ and we choose to have $M = 5$ minibathces, then each minibatch contains two datapoints. In particular we have $B_1 = (\mathbf{x}_1, \mathbf{x}_2), \cdots, B_5 = (\mathbf{x}_9, \mathbf{x}_{10})$. Note that if you choose $M = 1$ you have only a single batch with all datapoints and on the other extreme, you may choose $M = n$ resulting in a minibatch for each datapoint, i.e $B_k = \mathbf{x}_k$.

The idea is now to approximate the gradient by replacing the sum over all datapoints with a sum over the datapoints in one the minibatches picked at random in each gradient descent step

$$\begin{aligned} \nabla_\theta C(\theta) &= \sum_{i=1}^{n} \nabla_\theta c_i(\mathbf{x}_i, \theta) \\ &\rightarrow \sum_{i \in B_k}^{n} \nabla_\theta c_i(\mathbf{x}_i, \theta). \end{aligned} \tag{19}$$

Thus a gradient descent step now looks like

$$\theta_{j+1} = \theta_j - \gamma_j \sum_{i \in B_k}^{n} \nabla_\theta c_i(\mathbf{x}_i, \theta) \tag{20}$$

where k is picked at random with equal probability from the interval $[1, n/M]$. An iteration over the number of minibathces n/M is commonly referred to as an epoch. Thus it is typical to choose a number of epochs and for each epoch iterate over the number of minibatches.

Taking the gradient only on a subset of the data has two important benefits. First, it introduces randomness which decreases the chance that our opmization scheme gets stuck in a local minima. Second, if the size of the minibatches are small relative to the number of datapoints ($M < n$), the computation of the gradient is much cheaper since we sum over the datapoints in the k-th minibatch and not all n datapoints.

A natural question is when do we stop the search for a new minimum? One possibility is to compute the full gradient after a given number of epochs and check if the norm of the gradient is smaller than some threshold and stop if true. However, the condition that the gradient is zero is valid also for local minima, so this would only tell us that we are close to a local/global minimum. However, we could also evaluate the cost function at this point, store the result and continue the search. If the test kicks in at a later stage we can compare the values of the cost function and keep the θ that gave the lowest value.

Another approach is to let the step length γ_j depend on the number of epochs in such a way that it becomes very small after a reasonable time such that we do not move at all.

As an example, let $e = 0, 1, 2, 3, \cdots$ denote the current epoch and let $t_0, t_1 > 0$ be two fixed numbers. Furthermore, let $t = e \cdot m + i$ where m is the number of minibatches and $i = 0, \cdots, m-1$. Then the function

$$\gamma_j(t; t_0, t_1) = \frac{t_0}{t + t_1} \tag{21}$$

goes to zero as the number of epochs gets large. I.e. we start with a step length $\gamma_j(0; t_0, t_1) = t_0/t_1$ which decays in "time" t.

In this way we can fix the number of epochs, compute θ and evaluate the cost function at the end. Repeating the computation will give a different result since the scheme is random by design. Then we pick the final θ that gives the lowest value of the cost function.

MODEL SYSTEMS

In this work we apply the machine learning algorithms discussed to the one- and two-dimensional Ising model. Linear regression and neural networks are used to estimate the coupling constant of the one-dimensional Ising model.

The two-dimensional Ising model is known to show phase transition. In particular, these phases can be labeled as ordered or disordered. Thus it is interesting to investigate whether logistic regression or neural networks can be trained to to classify the phases.

The one-dimensional Ising model

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

The two-dimensional Ising model

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

RESULTS AND DISCUSSION

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetuer at, consectetuer sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.

Learning the one-dimensional Ising Hamiltonian

Morbi luctus, wisi viverra faucibus pretium, nibh est placerat odio, nec commodo wisi enim eget quam. Quisque libero justo, consectetuer a, feugiat vitae, porttitor eu, libero. Suspendisse sed mauris vitae elit sollicitudin malesuada. Maecenas ultricies eros sit amet ante. Ut venenatis velit. Maecenas sed mi eget dui varius euismod. Phasellus aliquet volutpat odio. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Pellentesque sit amet pede ac sem eleifend consectetuer. Nullam elementum, urna vel imperdiet sodales, elit ipsum pharetra ligula, ac pretium ante justo a nulla. Curabitur tristique arcu eu metus. Vestibulum lectus. Proin mauris. Proin eu nunc eu urna hendrerit faucibus. Aliquam auctor, pede consequat laoreet varius, eros tellus scelerisque quam, pellentesque hendrerit ipsum dolor sed augue. Nulla nec lacus.

Classifying phases of the two-dimensional Ising model

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi.

CONCLUSION

Sed feugiat. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Ut pellentesque augue sed urna. Vestibulum diam eros, fringilla et, consectetuer eu, nonummy id, sapien. Nullam at lectus. In sagittis ultrices mauris. Curabitur malesuada erat sit amet massa. Fusce blandit. Aliquam erat volutpat. Aliquam euismod. Aenean vel lectus. Nunc imperdiet justo nec dolor.