SVM Tutorial

SVM - Understanding the math - Unconstrained minimization

This is the **Part 4** of my series of tutorials about the math behind Support Vector Machines.
Today we are going to learn how to solve an unconstrained minimization problem.

If you did not read the previous articles, you might want to start the serie at the beginning by reading this article: an overview of Support Vector Machine.

About this Part

It took me a while to write this article because the subject is vast and assume a lot of prior knowledge. What should I explain and what should I skip was kind of a hard line to trace. After a while, I ended up with a large Part 4 which was too long to read. So I decided to split it. Welcome, Part 4, Part 5 and Part 6!

In this article try to make it as simple as possible for everybody. However, I cannot explain everything. I will assume that you know what **derivatives** and **partial derivatives** are. You are also expected to know what a matrix, the transpose of a matrix are and how to compute the determinant of a matrix.

During the last few months, I received a lot of comments and encouragements and several hundred people subscribed to be notified when this part is published. I wish to thank all of you, and I hope you will enjoy reading it.

Where we left.

In Part 3, we discovered that to maximize the margin we need to minimize the norm of $\mathbf{w}$.

It means we need to solve the following optimization problem:

$$\text{Minimize in } (\mathbf{w}, b)$$

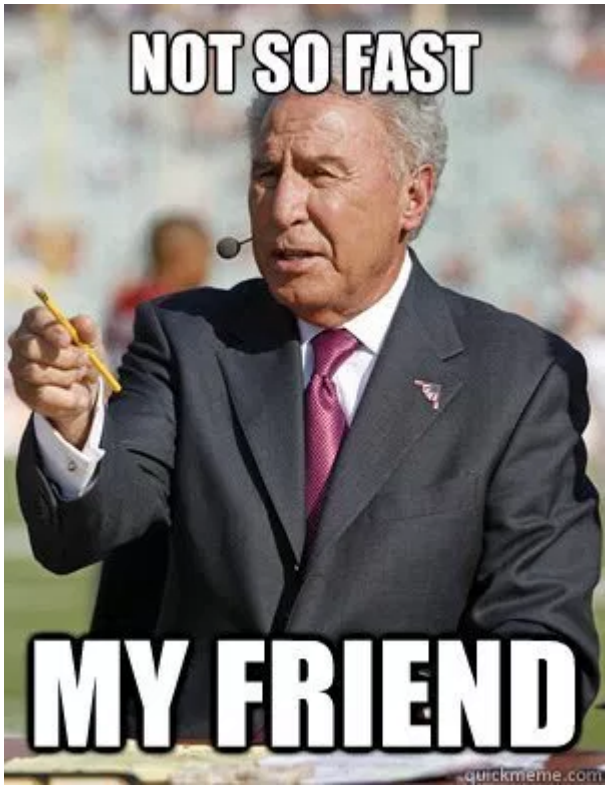
$$\|\mathbf{w}\|$$

$$\text{subject to } y_i(\mathbf{w} \cdot \mathbf{x_i} + b) \geq 1$$

(for any $i = 1, \ldots, n$)

The first thing to notice about this optimization problem is that it has **constraints**. They are defined by the line which begins with "subject to". You may think that there is only one constraint, but there is, in fact, n constraints. (this is because of the last line "for any"...)

"OK, How do I solve it? I have been waiting for this for one year !!!"



Before tackling such a complicated problem, let us start with a simpler one. We will first look at how to solve an unconstrained optimization problem, more specifically, we will study **unconstrained minimization**. That is the problem of finding which input makes a function return its minimum. (Note: in the SVM case, we wish to minimize the function computing the norm of $\mathbf{w}$, we could call it f and write it $f(\mathbf{w}) = \|\mathbf{w}\|$).

Unconstrained minimization

Let us consider a point $\mathbf{x}^*$ (you should read it "x star", we just add the star so that you know we are talking about a specific variable, and not about any $\mathbf{x}$).

How do we know if $\mathbf{x}^*$ is a local minimum of a function f? Well, it is pretty simple, we just need to apply the following theorem:

Theorem:

Let $f : \Omega \to \mathbb{R}$ be a continuously twice differentiable function at $\mathbf{x}^*$.

If $\mathbf{x}^*$ satisfies $\nabla f(\mathbf{x}^*) = 0$ and $\nabla^2 f(\mathbf{x}^*)$ is positive definite then $\mathbf{x}^*$ is a local minimum.

(Proof, at page 11)

The hard truth with such a theorem is that although being extremely concise, it is totally impossible to understand without some background information. What is $\nabla f(\mathbf{x}^*) = 0$? What is $\nabla^2 f(\mathbf{x}^*)$? What do we mean by positive definite?

Sometimes, we will be given more informations, and the previous theorem can also be rephrased like this :

Theorem (with more details):

If $\mathbf{x}^*$ satisfies:

1. f has a zero gradient at $\mathbf{x}^*$:

$$\nabla f(\mathbf{x}^*) = 0$$

and

2. the Hessian of f at $\mathbf{x}^*$ is positive definite:

$$\mathbf{z}^\intercal((\nabla^2 f(\mathbf{x}^*))\mathbf{z} > 0, \forall \mathbf{z} \in \mathbb{R}^n$$

where

$$\nabla^2 f(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

then $\mathbf{x}^*$ is a local minimum.

What does this all mean?

Let us examine this definition step by step.

Step 1:

Let $f : \Omega \to \mathbb{R}$ be a continuously twice differentiable function at $\mathbf{x}^*$.

First, we introduce a function which we call f, this function takes its values from a set Ω (omega) and returns a real number. There is a first difficulty here because we do not state what Ω is, but we will be able to guess it in the next line. This function f should be continuous and twice differentiable, or the rest of the definition will not be true.

Step 2:

$\mathbf{x}^*$ *is a local minimum of* $f(\mathbf{x})$ *if and only if:*

We want to find a value to give to f for it to produce its minimum. We simply name this value $\mathbf{x}^*$.

From the notation we can tell two things:

1. $\mathbf{x}^*$ is written in bold, so it is a vector. It means that f is a multivariate function.
2. As a result, the set Ω we saw earlier is the set from which we pick values to give to f. It means that the set Ω is a set of vectors and $\mathbf{x}^* \in \Omega$ ("x stars belongs to Omega")

Step 3:

f *has a zero gradient at* $\mathbf{x}^*$

This one is the first condition which must hold if we want $\mathbf{x}^*$ to be a local minimum of $f(\mathbf{x})$. We must check that the gradient of the function f at $\mathbf{x}^*$ is equal to zero. What is the gradient? Just think of it as a derivative on steroids.

> ***Definition***: *"the **gradient** is a generalization of the usual concept of derivative of a function in one dimension to a function in several dimensions" (Wikipedia)*

This definition gives us more pieces of information. A gradient is, in fact, the same thing as a derivative, but for functions like f which take vectors as input. That is why we wanted f to be a differentiable function in the first place, if it is not the case we cannot compute the gradient, and we are stuck.

In calculus, when we want to study a function, we often study the sign of its derivative. It allows you to determine if the function is increasing or decreasing and to identify minimum and maximum. By setting the derivative to zero, we can find the "critical points" of the function at which it reaches a maximum or a minimum. (You can read this excellent explanation if you want to refresh your memory). When we work with functions having more variable, we need to set each partial derivative to zero.

It turns out, the gradient of a function is a vector containing each of its partial derivatives. By studying the sign of the gradient, we can gather important pieces of information about the function. In this case, checking if the gradient equals zero for $\mathbf{x}^*$ allow us to determine if $\mathbf{x}^*$ is a critical point (and that the function f possibly has a minimum at this point). (Note: Checking if the gradient equals zero at a point means checking that each partial derivative equals zero for this point)

The gradient of a function is denoted by the symbol ∇ (nabla).
The line

$$\nabla f(\mathbf{x}^*) = 0$$

is just a repetition of "f has a zero gradient at $\mathbf{x}^*$" in mathematical notation.

For a vector $\mathbf{x}^*(x_1, x_2, x_3)$, $\nabla f(\mathbf{x}^*) = 0$ means:

$$\frac{\partial f}{\partial x_1}(\mathbf{x}^*) = 0$$

$$\frac{\partial f}{\partial x_2}(\mathbf{x}^*) = 0$$

$$\frac{\partial f}{\partial x_3}(\mathbf{x}^*) = 0$$

Step 4:

> *the Hessian of f at $\mathbf{x}^*$ is positive definite*

That is where most people get lost. This single sentence requires a lot of backgrounds. You need to know:

1. that the Hessian is a matrix of second-order partial derivatives
2. how we can tell if a matrix is positive definite

The Hessian matrix

The Hessian is a matrix, and we give it a name. We could call it H but instead we call it $\nabla^2 f(\mathbf{x})$ which is more explicit. We keep the symbol ∇ used for the gradient, and add a ^{2} to denote we the fact that this time we are talking about second-order partial derivative. Then we specify the name of the function (f) from which we will compute these derivates. By writing $f(\mathbf{x})$ we know that f takes a vector $\mathbf{x}$ as input and that the Hessian is computed for a given $\mathbf{x}$.

To sum up, we need to compute a matrix called the Hessian matrix for $\mathbf{x}^*$.
So we take the function f, we take the value of $\mathbf{x}^*$ and we compute the value for each cell of the matrix using the following formula:

$$\nabla^2 f(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

Eventually we get the Hessian matrix and it contains all the numbers we have computed.

Let us look at the definition to see if we understand it well:

> ***Definition**: In mathematics, the **Hessian matrix** or **Hessian** is a square matrix of second-order partial derivatives of a scalar-valued function.*
> ***It describes the local curvature of a function of many variables**. (Wikipedia)*

(Note: A scalar valued function is a function that takes one or more values but returns a single value. In our case f is a scalar valued function.)

Positive definite

Now that we have the Hessian matrix, we want to know if it is positive definite at $\mathbf{x}^*$.

> ***Definition**: A symmetric matrix A is called **positive definite** if $\mathbf{x}^\intercal A\mathbf{x} > 0$, for all $\mathbf{x} \in \mathbb{R}^n$. (Source)*

This time, we note that once again we were given the definition in the first place. It was just a little bit harder to read because of our notational choice. If we replace A by $\nabla^2 f(\mathbf{x}^*)$ and $\mathbf{x}$ by $\mathbf{z}$ we get exactly the formula written in the part 2. of the detailed theorem:

$$\mathbf{z}^\intercal((\nabla^2 f(\mathbf{x}^*))\mathbf{z} > 0, \forall \mathbf{z} \in \mathbb{R}^n$$

The problem with this definition is that it is talking about a symmetric matrix. A **symmetric matrix** is a square matrix this is equal to its transpose.

The Hessian matrix is square, but is it symmetric?

Luckily for us yes!

> *"if the second derivatives of f are all continuous in a neighborhood D, then the Hessian of f is a symmetric matrix throughout D" (Wikipedia)*

But even with the definition, we still don't know how to check that the Hessian is positive definite. That is because the formula $\mathbf{z}^\intercal((\nabla^2 f(\mathbf{x}^*))\mathbf{z} \geq 0$ is for all $\mathbf{z}$ in $\mathbb{R}^n$.

We can't try this formula for all $\mathbf{z}$ in $\mathbb{R}^n$!
That is why we will use the following theorem:

Theorem:
The following statements are equivalent:

- The symmetric matrix A is positive definite.
- All eigenvalues of A are positive.
- All the leading principal minors of A are positive.
- There exists nonsingular square matrix B such that $A = B^\intercal B$ (Source)

So we have three ways of checking that a matrix is positive definite:

- By computing its eigenvalues and checking they are positive.
- By computing its leading principal minors and checking they are positive.
- By finding a nonsingular square matrix B such that $A = B^\intercal B$.

Let's pick the second method and look at it in more details.

Computing the leading principal minors

Minors

To compute the minor M_{ij} of a matrix we remove the i^{th} line and the j^{th} column, and compute the determinant of the remaining matrix.

Example:
Let us consider the following 3 by 3 matrix:

$$\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}$$

To compute the minor M_{12} of this matrix we remove the line number 1 and the column number 2. We get:

$$\begin{pmatrix} \square & \square & \square \\ d & \square & f \\ g & \square & i \end{pmatrix}$$

so we compute the determinant of:

$$\begin{pmatrix} d & f \\ g & i \end{pmatrix}$$

which is : $di - fg$

Principal minors

A minor M_{ij} is called a **principal minor** when $i = j$.

For our 3x3 matrix, the principal minors are :

- $M_{11} = ei - fh$,
- $M_{22} = ai - cg$
- $M_{33} = ae - bd$

But that is not all ! Indeed, minors also have what we call an order.

Definition:
A minor of A of order k is principal if it is obtained by deleting $n - k$ rows and the $n - k$ columns with the same numbers. (Source)

In our previous example, the matrix is 3×3 so $n = 3$ and we deleted 1 line, so we computed principal minors of order 2.

There are $\binom{n}{k}$ □ principal minors of order k, and we write Δk for any of the principal minors of order k.

To sum-up:

Δ_0: does not exist because if we remove three lines and three columns we have deleted our matrix!

Δ_1: We delete (3-1) = 2 lines and 2 columns with the same number.
So we remove lines 1 and 2 and column 1 and 2.

$$\begin{pmatrix} □ & □ & □ \\ □ & □ & □ \\ □ & □ & i \end{pmatrix}$$

It means that one of the principal minors of order 1 is i. Let us find the others:

We delete lines 2 and 3 and column 2 and 3 and we get a.
We delete lines 1 and 3 and column 1 and 3 and we get e

Δ_2: is what we have seen before:

- $M_{11} = ei - fh$,
- $M_{22} = ai - cg$
- $M_{33} = ae - bd$

Δ_3: We delete nothing. So it is the determinant of the matrix : $aei + bfg + cdh - ceg - bdi - afh$.

Leading principal minor

Definition:
The leading principal minor of A of order k is the minor of order k obtained by deleting the last $n - k$ rows and columns.

So it turns out leading principal minors are simpler to get. If we write D_k for the leading principal minor of order k we find that:

$D_1 = a$ (we deleted the last two lines and column)

$D_2 = ae - bd$ (we removed the last line and the last column)

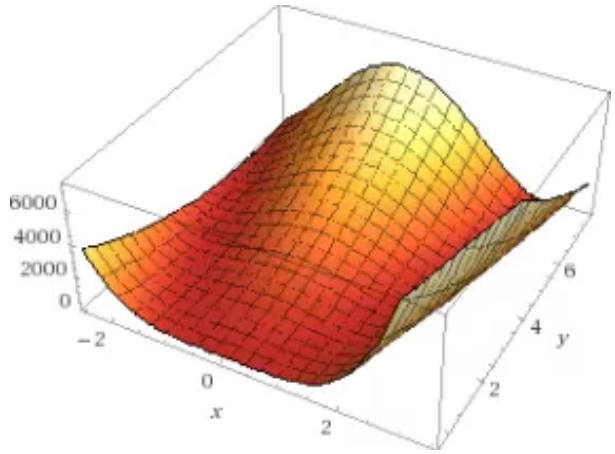
$D_3 = aei + bfg + cdh - ceg - bdi - afh$

Now that we can compute all the leading principal minors of a matrix, we can compute them for the Hessian matrix at $\mathbf{x}^*$ and if they are all positive, we will know that the matrix is positive definite.

We now have fully examined what we have to know, and you should be able to understand how to solve an unconstrained minimization problem. Let us check that everything is clear with an example.

Example:

In this example we will try to find the minimum of the function: $f(x,y) = (2-x)^2 + 100(y-x^2)^2$ which is known as the Rosenbrock's banana function.



The Rosenbrock function for a = 2 and b = 100

First, we will search for which points its gradient $\nabla f(x,y)$ equals zero.

$$\nabla f(x,y) = \begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{pmatrix}$$

So we compute the partial derivatives and we find:

$$\frac{\partial f}{\partial x} = 2(200x^3 - 200xy + x - 2)$$

$$\frac{\partial f}{\partial y} = 200(y - x^2)$$

(Tip: if you want to check your calculation you can use Wolfram alpha)

Our goal is to find when they are both at zero, so we need to solve the following system of equations:

$$2(200x^3 - 200xy + x - 2) = 0 \tag{1}$$
$$200(y - x^2) = 0 \tag{2}$$

We distribute to get:

$$400x^3 - 400xy + 2x - 4 = 0 \tag{3}$$
$$200y - 200x^2 = 0 \tag{4}$$

We multiply (2) by $2x$ to obtain:

$$400xy - 400x^3 = 0 \tag{5}$$

We now add (3) and (5) to get:

$$400x^3 - 400xy + 2x - 4 + 400xy - 400x^3 = 0 \tag{6}$$

which simplifies into:

$$2x - 4 = 0$$

$$x = 2$$

We substitute x in (4)

$$200y - 200 \times 2^2 = 0$$

$$200y - 800 = 0$$

$$y = \frac{800}{200}$$

$$y = 4$$

It looks like we have found the point $(x, y) = (2, 4)$ for which $\nabla f(x, y) = 0$. But is this a minimum?

The Hessian matrix is :

$$\nabla^2 f(x,y) = \begin{pmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial xy} \\ \frac{\partial^2 f}{\partial yx} & \frac{\partial^2 f}{\partial y^2} \end{pmatrix}$$

$$\frac{\partial^2 f}{\partial x^2} = 1200x^2 - 400y + 2$$

$$\frac{\partial^2 f}{\partial xy} = -400x$$

$$\frac{\partial^2 f}{\partial yx} = -400x$$

$$\frac{\partial^2 f}{\partial y^2} = 200$$

Let us now compute the Hessian for $(x,y) = (2,4)$

$$\nabla^2 f(x,y) = \begin{pmatrix} 3202 & -800 \\ -800 & 200 \end{pmatrix}$$

The matrix is symetric, we can check its leading principal minors:

Minors of rang 1:

If we remove the last line and last column the minor M_{11} is 3202.

Minor of rang 2:

This is the determinant of the Hessian:

$$3202 \times 200 - (-800) \times (-800) = 400$$

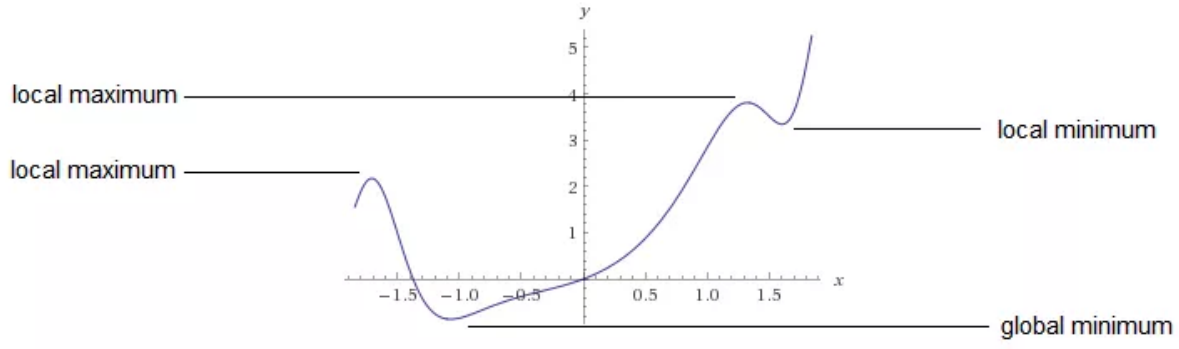
All the leading principal minors of the Hessian are positives. It means that the Hessian is positive definite.

The two conditions we needed are met, and we can say that the point $(2, 4)$ is a local minimum.

LOCAL minimum?

A point is called a **local minimum** when it is the smallest value within a range. More formally:

Given a function f defined on a domain X, a point x^* is said to be a local minimum if there exists some $\epsilon > 0$ such that $f(x^*) \leq f(x)$ for all x in X within distance ϵ of x^*. This is illustrated in the figure below:



A **global minimum**, however, is true for the whole domain of the function:

Given a function f defined on a domain X, a point x^* is said to be a **global** minimum if $f(x^*) \leq f(x) \quad \text{for all } x \text{ in } X$

So all our hard work was just to find a local minimum, but in real life, we often want to find the global minimum...

How can we find a global minimum?

There is one simple way to find the global minimum:

1. Find all the local minima
2. Take the smallest one; it is the global minimum.

Another approach is to study the function we are trying to minimize. If this function is **convex**, then we are sure its local minimum is a global minimum.

Conclusion

We discovered that finding the minimum of a function is not so simple, and it was not even a global minimum. However, some functions, called convex functions are easier to work with. What is a convex function? Read the Part 5 of this tutorial series to find out! Thanks for reading.

Alexandre KOWALCZYK

I am passionate about machine learning and Support Vector Machine. I like to explain things simply to share my knowledge with people from around the world. If you wish you can add me to linkedin, I like to connect with my readers.

This entry was posted in Mathematics, SVM Tutorial and tagged Hessian, matrix, minimization, unconstrained on September 11, 2016 [https://www.svm-tutorial.com/2016/09/unconstrained-minimization/] .

31 thoughts on “SVM - Understanding the math - Unconstrained minimization”

kecai wu

September 19, 2016 at 1:23 pm

a point x* is said to be a global minimum if f(x*)>=f(x) for all x in X? This x* is a global maximum!

Alexandre KOWALCZYK Post author

September 19, 2016 at 10:23 pm

Thank you. I corrected the typo.

Ravindra M

September 21, 2016 at 7:50 am

Thank you for writing this article. Explaining the math along the way intuitively and in detail.

Eric L

September 26, 2016 at 2:21 pm

Thank you for the attention you pay for explaining every details. It's really helpfull. By the way, I'm not sure but is it normal that f partial derivative relative to x is exactly equal to f(x,y) ?
(2-x)^2+100(y-x^2)^2

Alexandre KOWALCZYK Post author

September 26, 2016 at 10:08 pm

Indeed it was a typo. I fixed it. Thanks a lot.

LuyuCHEN

November 4, 2016 at 4:22 am

Thank you for your excellent explanation about support vector machine! It is so helpful!!!!

Nick

November 15, 2016 at 9:35 pm

It's the best explanation about svm which I ever was seen!

lcmosasohi

November 17, 2016 at 7:36 am

Sorry I do not understand why we compute M22 in the given example if "minor of order kk obtained by deleting the last n−kn−k rows and columns." It seemed it just needed to compute M11 when delete the last 1 raw and columb ...

Alexandre KOWALCZYK Post author

November 18, 2016 at 4:28 pm

Hello. You are right, we do not need to calculate M22 as it is not a leading principal minor. I updated the article to avoid the confusion. Thank you for your comment 🙂

Yixin Chen

November 18, 2016 at 9:32 pm

Hi, thanks for the article. It's pretty clear and thorough. Just on confusion I thought in your example of Rosenbrock's banana function, should M11 be 4801 instead of 200 because your try to remove the last line and column. Actually I think this is the same question as lcmosasohi put. Thanks.

Alexandre KOWALCZYK Post author

November 19, 2016 at 1:24 pm

Hello Yixin,

Yes indeed, you are correct. I updated the article.

Thank your for your comment.

Igor Petrovski

November 30, 2016 at 9:06 pm

Man, this is amazing!
I am doing my master studies at ETH Zürich and this helps me a lot!
Do you have any other tutorials on some other models (Linear Regression, Logistic Regression, etc) ?

Cheers and thank you!

Alexandre KOWALCZYK Post author

December 4, 2016 at 6:39 pm

Hello Igor. Thank you for your comment. For the moment I only have tutorials about SVM. 🙂

gsk

January 7, 2017 at 7:55 pm

Dear Alex, you're doing great job here.Please continue the same.

If possible, Please make tutorials on Math behind on other machine learning algorithms as well. You're going to be life savior to many.

Thank you in advance.

Srinivas

December 13, 2016 at 5:49 am

Sir, when your book will be available? Are you also covering the kernels in that? eager to see the book.

张翔

February 23, 2017 at 1:25 pm

I hope to read more tutorial about Random forest, Logistic Regression, boosting etc. if you are free.

张翔

February 23, 2017 at 1:23 pm

Thanks very much!
This is the best SVM tutorial I met ever!!!

Felipe

February 23, 2017 at 9:00 pm

First of all, thank you very much for this article.
I have a question about hessian matrix, why did you say that the first element of the matrix is 4801? Shouldn't it be $1200 \times 2^2 - 400 \times 4 + 2 = 3202$?

Alexandre KOWALCZYK Post author

February 23, 2017 at 9:27 pm

Yes indeed, I corrected the formula and the value. Thanks a lot!

Felipe

February 23, 2017 at 10:12 pm

I'm happy to help. Your tutorial is the best. Thanks again.

Minseok

February 24, 2017 at 3:41 am

I study neural network in korea university student, recently, I'm happy to study, because I read your lecture to SVM. You are best teacher for me. I hope that you lecture for Principal Component Analisis. thank you so much. if you will write to lecture, i must read to your lecture.

Nitin Bansal

February 25, 2017 at 12:04 am

Best Article on SVM. Thank You!

Ravindra

March 26, 2017 at 7:11 am

Thanks for excellent explanation

Yi Longhao

April 15, 2017 at 4:58 am

Thank you very much!

Paul Navin

May 25, 2017 at 3:48 pm

This is the most intuitive and lucid explanation of math behind statistical model I've ever seen. Thank's a lot!

Jayesh

July 8, 2017 at 2:17 pm

Hi Alex,

Great write-up indeed. Helped me a lot.
Are you planning to take-up "regularizer terms" to avoid overfitting while maximizing margins?

Alexandre KOWALCZYK Post author

July 14, 2017 at 8:40 pm

Not in this series. I talk about soft margin SVM in my ebook though.

Mirek

August 14, 2017 at 5:48 pm

Hello Alexandre,

thanks a lot for this tutorial.
I have one question regarding Face Recognition with Support Vector Machines. What is in this case the support vector (Xi): one pixel of the picture or the whole picture as a big matrix with all pixels? In the tutorials will often be used an example with 2D dots. Could you please explain how it works? Thanks!

regards

mirek

Alexandre KOWALCZYK Post author

August 14, 2017 at 11:44 pm

In this case Xi represent the whole picture as a vector with all the pixel. X is a matrix containing all the images (each image is a row).

Mirek

August 15, 2017 at 9:36 am

Thanks a lot, I saw that for this case can be useful a Principal Component Analysis (PCA) in order to reduce number of pixels (dimention of feature vector).

Kevin

November 7, 2017 at 6:08 am

Great great thanks to you, it is hard to understand all the slide in the lecture, u really make all the SVMs technical terms and concept much more straightforward!!! Thanks again

---