Week 5: Support Vector Machines and Expectation-Maximization

March 1, 2019

1 Support Vector Machines

1.1 Support Vector Loss

The basic optimization objective for Support Vector Machines is

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|\boldsymbol{w}\| + C\sum_i p_i \\ \text{such that} \quad & y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geqslant 1 - p_i \\ \text{and} \quad & p_i \geqslant 0 \end{aligned}$$

Exercise 1

What does i index? How many terms does the sum have, and how many constraints are there?

It iterates over the data. There are as many terms as there are instances in the data. There are two constraints per instance.

What is the value of y_i in this expression? What is its function?

y_i is 1 for positive examples, and -1 for negative examples. It allows us to rewrite the constraints

$$\begin{aligned} \boldsymbol{w}^T\boldsymbol{x}_i + b \geqslant 1 \text{ if } \boldsymbol{x}_i \text{ is positive} \\ \boldsymbol{w}^T\boldsymbol{x}_i + b \leqslant -1 \text{ if } \boldsymbol{x}_i \text{ is negative} \end{aligned}$$

to a single constraint.

Exercise 2

There are two common ways to rewrite this expression before implementing it. What are they (in general terms) and what are their benefits?

The first option is to rewrite everything in terms of $\boldsymbol{w}$ and b in order to get rid of the constraints. This is useful when we want to use the SVM as the last layer in a neural network. Without constraints, we are free to use basic backpropagation.

The second option is to use Lagrange multipliers to get rid of $\boldsymbol{w}$ and rewrite everything in terms of the multipliers. This exresses the solution purely in terms of the support vectors.

The main benefit is that the whole algorithm can be written in terms of the dot products of pairs of instances. This means we do not need to see the actual feature vectors to compute the support vectors, only the dot products. This allows us to apply the *kernel trick*.

1.2 Lagrange multipliers

Lagrange multipliers are a useful trick to know. We'll practice them briefly on a small problem, so that you understand the principle. We will only use an *equality* constraint.

We have the following optimization problem:

$$\begin{aligned} \text{minize } f(a, b) &= a^2 + 2b^2 \\ \text{such that } a^2 &= -b^2 + 1 \end{aligned}$$

Exercise 3

The first step is to rewrite the constraint so that the right side is equal to zero. Do so.

$a^2 + b^2 - 1 = 0$ What does the constraint say about the allowed inputs (what shape do the allowed inputs make in the (a, b)-plane)?

The solutions are constrained to a circle, centered on the origin, with radius 1 (the so-called bi-unit circle).

We now define a function $L(a, b, \lambda) = f(a, b) + \lambda G$, where G is the left hand side of the constraint equal to zero (how much any given a and b violate the constraint).[1]

Exercise 4

Write out $L(a, b, \lambda)$ for our problem. $L(a, b, \lambda) = a^2 + 2b^2 + \lambda(a^2 + b^2 - 1)$

We take the derivative of L with respect to each of its *three* parameters, and set these equal to zero.

Exercise 5

Fill in the blanks

$$\begin{aligned}
\frac{\partial L}{\partial a} &= \frac{\partial(a^2 + 2b^2 + \lambda a^2 + \lambda b^2 - \lambda)}{\partial a} \\
&= 2a + 2\lambda a \\
\frac{\partial L}{\partial b} &= 4b + 2\lambda b \\
\frac{\partial L}{\partial \lambda} &= a^2 + b^2 - 1
\end{aligned}$$

$$a(2 + 2\lambda) = 0 \tag{1}$$
$$b(4 + 2\lambda) = 0 \tag{2}$$
$$a^2 + b^2 = 1 \tag{3}$$

Note that the last line recovers the original constraint. We now have three equations with three unknowns, so we can solve for a and b. From the shape of the function (it's symmetric in both the a and b axes), we should expect at least two solutions.

We can get these from the above equations by noting that if a and b are both nonzero, we can derive a contradiction. Thus either a or b must be zero.

Exercise 6

Give the solutions for both cases (remember that $x^2 = 1$ has *two* solutions).

From line (3), above we see that if $a = 0$, then $b^2 = 1$ and vice versa.

[1] For plain Lagrange multipliers, where the constraints are all equalities, we can either add or subtract the term containing the constraint. For inequality constrains, it depends on whether we are maximizing or minimizing.

This gives us

$$\begin{aligned} a = 0, b = 1 \\ a = 0, b = -1 \\ a = 1, b = 0 \\ a = -1, b = 0 \end{aligned}$$

as extrema. Filling these in, we find that the last two lines minimize the function. (The first two are maxima, as can be seen by clicking the Wolfram Alpha link below.

Happily, Wolfram Alpha agrees with us (and provides some informative plots).

1.3 The kernel trick

The feature space of k is a projection of point $\mathbf{a}$ to point $\mathbf{a}'$ such that

$$k(\mathbf{a}, \mathbf{b}) = \mathbf{a}'^T \mathbf{b}' \; .$$

We have a dataset with two features Let $\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$ and $\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$. We define the *kernel*

$$k_1(\mathbf{a}, \mathbf{b}) = (\mathbf{a}^T \mathbf{b})^2 \; .$$

Exercise 7

Show that the feature space defined by$k_!$ is

$$\begin{pmatrix} {x_1}^2 \\ \sqrt{2} x_1 x_2 \\ {x_2}^2 \end{pmatrix} \; .$$

Hint: start by writing out the definition as a scalar function. See if your can re-arrange this back into a dot procut of two other vectors.

Starting from the definition of k:

$$
\begin{aligned}
k(\mathbf{a}, \mathbf{b}) &= \left(\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}^T \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \right)^2 \\
&= (a_1 b_1 + a_2 b_2)^2 \\
&= a_1 b_1 a_1 b_1 + 2 a_2 b_2 a_1 b_1 + a_2 b_2 a_2 b_2 \\
&= a_1 a_1 \cdot b_1 b_1 + 2 \cdot a_1 a_2 \cdot b_1 b_2 + a_2 a_2 \cdot b_2 b_2 \\
&= \begin{pmatrix} a_1 a_1 \\ \sqrt{2} a_1 a_2 \\ a_2 a_2 \end{pmatrix}^T \begin{pmatrix} b_1 b_1 \\ \sqrt{2} b_1 b_2 \\ b_2 b_2 \end{pmatrix}
\end{aligned}
$$

Exercise 8

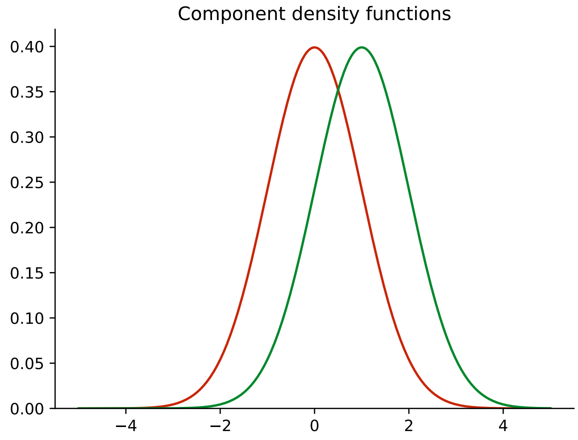
What is the feature space for the kernel

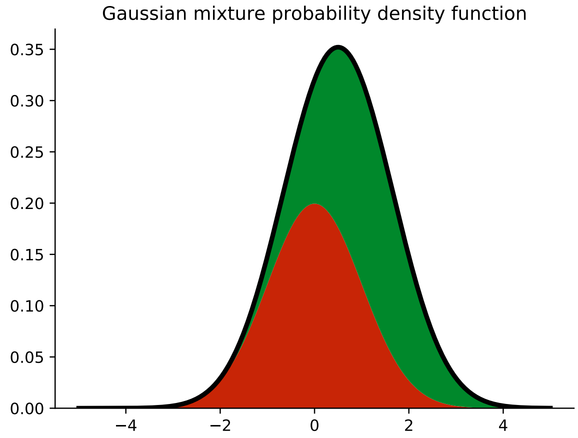
$$
k_2(\mathbf{a}, \mathbf{b}) = \left(\mathbf{a}^T \mathbf{b} + 1 \right)^2 \quad ?
$$

$$
\begin{aligned}
k(\mathbf{a}, \mathbf{b}) &= \left(\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}^T \begin{pmatrix} b_1 \\ b2 \end{pmatrix} + 1 \right)^2 \\
&= (a_1 b_1 + a_2 b_2)^2 + 2(a_1 b_1 + a_2 b_2) + 1 \\
&= a_1 b_1 a_1 b_1 + 2 a_1 b_1 a_2 b_2 + a_2 b_2 a_2 b_2 + 2 a_1 b_1 + 2 a_2 b_2 + 1 \\
&= \begin{pmatrix} 1 \\ \sqrt{2} a_1 \\ \sqrt{2} a_2 \\ {a_1}^2 \\ \sqrt{2} a_1 a_2 \\ {a_2}^2 \end{pmatrix}^T \begin{pmatrix} 1 \\ \sqrt{2} b_1 \\ \sqrt{2} b_2 \\ {b_1}^2 \\ \sqrt{2} b_1 b_2 \\ {b_2}^2 \end{pmatrix}
\end{aligned}
$$

2 Expectation Maximization

Assume we have a Gaussian Mixture Model in one dimension with two components: $N(0, 1)$ and $N(1, 1)$. The weights w_1 and w_2 of the components are equal.


Component density functions


Gaussian mixture probability density function

Exercise 9

Compute the probability density of the point 0, under the Gaussian Mixture.

$$\begin{aligned} p(0) &= \frac{1}{2}\mathrm{N}(0,1) + \frac{1}{2}\mathrm{N}(1,1) \\ &= \frac{1}{2\sqrt{2\pi}}\exp(0) + \frac{1}{2\sqrt{2\pi}}\exp\left(-\frac{1}{2}\right) \\ &= \frac{1}{2}\frac{1}{\sqrt{2\pi}} + \frac{1}{2}\frac{1}{\sqrt{2\pi e}} \approx 0.32 \end{aligned}$$

Exercise 10

Under the EM algorithm, what responsibility is assigned to each component for the point 0?

To compute the responsibility, we compute the probability of the component z given the point x: $p(z \mid x)$, and normalize over all components. We get for component 1:

$$\frac{\frac{1}{2\sqrt{2\pi}}\exp(0)}{\frac{1}{2\sqrt{2\pi}}\exp(0)+\frac{1}{2\sqrt{2\pi}}\exp\left(-\frac{1}{2}\right)}=\frac{1}{1+\frac{1}{\sqrt{e}}}\approx 0.62$$

and for component 2:

$$\frac{\frac{1}{2\sqrt{2\pi}}\exp\left(-\frac{1}{2}\right)}{\frac{1}{2\sqrt{2\pi}}\exp(0)+\frac{1}{2\sqrt{2\pi}}\exp\left(-\frac{1}{2}\right)}=\frac{\frac{1}{\sqrt{e}}}{1+\frac{1}{\sqrt{e}}}\approx 0.37$$

Note: Using Bayes' rule, this translates to $p(z \mid x) = \frac{p(x|z)p(z)}{p(x)}$. The denominator is the sum computed in the previous exercise, and the responsibilities are the proportions of each term to the total.[2]

This is no accident, it is essentially what Bayes' rule tells us: to compute $p(z \mid x)$, we find $p(x)$ by marginalizing Z out of $p(X = x, Z)$. This gives us a big sum, with one term for each z. The proportion of this term to the total is $p(z \mid x)$.