Machine Learning B
Home Assignment 1

Jakob Schauser, pwn274

November 2021

1 Convex Functions

(a)
The easiest way to figure out whether this function is convex or not is to look at the hessian. As $f(x)$ is a function of a single variable, this amounts to taking the double derivative[1]:

$$f''(x) = (\log(x) + x/x + (-1)\log(1-x) - (1-x)/(1-x))' = 1/x + 1/(1-x)$$

As x is in the interval $(0, 1)$ I can see that this is always positive, with a minimum of $f(x) = 4$ at $x = 1/2$ going towards infinity at the boundaries. As $f''(x)$ is strictly positive, $f(x)$ is strictly convex. Checking whether it is also strongly convex, I add the $\alpha/2x^2$-term, which clearly just corresponds to adding a $-\alpha$ term in the Hessian. As I require $f''(x) \geq 0$ this means $f(x)$ is strongly convex for $\alpha < 4$, i.e. 4-strongly convex.

(b)
Again, taking the second derivative, I can see that it is always positive albeit going towards 0 for $x \to \infty$:

$$\nabla^2 = \frac{1}{4(1+x)^{3/2}} \tag{1}$$

As it is always positive, we once again have a strictly convex function on our hands. This time, there is no possible α to always have $\nabla^2 \geq 0$, so the function is not strongly convex.

(c)
Using Wolfram-Alpha (which I am pretty sure is allowed), I can see that the Hessian is symmetric and only has real positive eigenvalues which is equivalent to it being positive-definite. This means, $f(\mathbf{x})$ is strictly convex. I am unsure how to check for strongness, but

[1] Another way would be to test the definition $f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y)$

[2] This is both an exponent and a footnote. I am sorry for swapping between notation, but f" and ∇^2 means the same thing for most of these functions

as both variables show up in the equation proportional to x^2 and I have shown it to be strictly convex, I imagine this to be the case.

(d)

All norms are described in the slides on p. 11/45 as being convex. As the function is linear, it is easy to see the convexity is neither strict nor strong.

(e)

Here, $f(x)$ is built from parts we have seen before. It can easily be surmised, that $f(x)$ is convex for $\lambda = 0$ as this means the function is a simple linear[3], strictly convex for $\lambda > 0$ as we have a parabola, and α-strongly convex for $\alpha < 2\lambda$. If $\lambda < 0$ the function is strictly concave.

2 KKT Conditions

2.1

Here, the main problem lies in transforming the x_i-bounds to the standard form. As the set in which x_i lies is convex, I can simply change the upper and lower bounds into inequalities.

$$\sum_{i=1}^{d} x_i \leq C \leftrightarrow \sum_{i=1}^{d} x_i - C \leq 0 \tag{2}$$

$$x_i \leq [m_i, M_i] \rightarrow m_i - x_i \leq 0 \;\; \& \;\; x_i - M_i \leq 0 \tag{3}$$

This gives the final formulation:

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{d} \frac{1}{x_i+\alpha_i} \quad x \in M \\ \text{subject to} & \sum_{i=1}^{d} x_i - C \leq 0 \\ & m_i - x_i \leq 0 \\ & x_i - M_i \leq 0 \quad i = 1, \ldots, d \end{array} \tag{4}$$

2.2

Look at any x_i, I can see that the second derivative of the function is $2/(a+x)^3$. As both a and x are positive, this expression is also positive, which corresponds to the second derivative of the sum also being positive.

From this, I can gather that the whole function is convex. As it is also clear, that the set is convex (it is 1-dimensional and has no holes) for every x_i, the problem is a convex optimization problem.

[3] or affine, to be excact

2.3

Defining the Lagrangian is now as simple as putting it in the form as found in the notes, where we are careful to give each constraint their individual lambda:

$$\mathcal{L} = \sum_{i=1}^{d} \frac{1}{x_i + \alpha_i} + \lambda_0 \sum_{i=1}^{d} (x_j - C) + \sum_{i=1}^{d} \lambda_{i+1}(m_i - x_i) + \sum_{i=1}^{d} \lambda_{d+i+1}(x_i - M_i) \tag{5}$$

2.4

We know that both the initial function and its constraints are convex functions that can be put in the form of the KKT. I am unsure what more to write, as 2.1 basically already asked us to think of the problem in the standard form with KKT-conditions and I have shown that f and g_i are convex functions.

2.5

The function:

$$D(\lambda) = \min_x L(x, \lambda) = \min_x \sum_{i=1}^{d} \frac{1}{x_i + \alpha_i} + \lambda_0 \sum_{i=1}^{d} (x_j - C) + \sum_{i=1}^{d} \lambda_{i+1}(m_i - x_i) + \sum_{i=1}^{d} \lambda_{d+i+1}(x_i - M_i) \tag{6}$$

The problem is thus maximizing this dual function, which we notate as follows::

$$\sup_{\lambda \geq 0, \mu} \left(\min_x \sum_{i=1}^{d} \frac{1}{x_i + \alpha_i} + \lambda_0 \sum_{i=1}^{d} (x_j - C) + \sum_{i=1}^{d} \lambda_{i+1}(m_i - x_i) + \sum_{i=1}^{d} \lambda_{d+i+1}(x_i - M_i) \right) \tag{7}$$

3 Perceptron as a Subgradient Descent Algorithm

3.1

As w^* is defined as the w with the minimal norm, $w : \|w\| \leq \|w^*\|$ is all the w with a norm equal to $\|w^*\|$. As we know f is maximizing and $y \geq 1$, we can surmise that $f(w : \|w\| = \|w^*\|) = 0$.

Why this is enough information to see, that all $f(w) < 1$ can cleanly separate is probably trivial to see, but I cannot seem remember it from the lecture or find it in the notes.

3.2

We know the requirement, that any $y \in \mathcal{X}$.

$$f(y) \geq f(x) + g^\top (y - x) \tag{8}$$

While the visual representation of this property is a lot more intuitive, the preceding requirement can be used to calculate a given subgradient. Finding a

3.3

Here I am also unsure how to work out the problem : (

4 Kernels

4.1 Distance in feature space

I simply get to work using the standard norm:

$$\|\Phi(x) - \Phi(z)\| = \sqrt{\langle \Phi(x) - \Phi(z), \Phi(x) - \Phi(z) \rangle} =$$
$$\sqrt{\langle k(x,\cdot) - k(z,\cdot), k(x,\cdot) - k(z,\cdot) \rangle} = \sqrt{(k(x,\cdot) - k(z,\cdot))^2} =$$
$$\sqrt{(k(x,x) - k(z,x) - k(x,z) + k(z,z))} = \sqrt{k(x,x) - 2k(x,z) + k(z,z)}$$

Here I used the standard dot-product, the fact that $k(a,\cdot)k(b,\cdot) = k(a,b)$ and the fact that $k(b,a) = k(a,b)$.

4.2 Sum of kernels

We start with:

$$k(x,z) = a \cdot k_1(x,z) + b \cdot k_2(x,z) \tag{9}$$

we know $x^T K_1 x \geq 0$ and $x^T K_2 x \geq 0$ for all x, this means that their sum also holds:

$$0 \leq x^T K_1 x + x^T K_2 x \leq a x^T K_1 x + b x^T K_2 x \tag{10}$$

(for $a, b > 0$) We can now use the distributive quality of matricies and the commuting property of real numbers to see:

$$0 \leq a x^T K_1 x + a x^T K_2 x = x^T a K_1 x + x^T b K_2 x = x^T (a K_1 + b K_2) x \tag{11}$$

As this is the definition for a positive definite matrix defined by the function $k(x,z) = a \cdot k_1(x,z) + b \cdot k_2(x,z)$, I can say QED

4.3 Rank of Gram matrix

As the matrix need only be symmetry and with shape $d \times d$, a trivial maximal rank is d. But given only $m < d$ input patterns, the requirement for the $x^T K x$ loosens and some columns of K are free to become linear combinations of one-another leaving a rank of m. This means, the rank of the Gram matrix will be $min(m, d)$.