Homework 4
Ada Huang
CSCE 633 - 600
April 7, 2023

1 Convex Functions

log-sum-exponential function convexity

Proof. To prove that the log-sum-exp function is convex, we will use the second-order condition of convexity, which states that a function is convex if and only if its Hessian matrix is positive semi-definite.

The log-sum-exp function is defined as:

$$f(\mathbf{x}) = \log\left(\sum_{i=1}^{n} \exp(x_i)\right)$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_n)^\top$.

Let's compute the first and second derivatives of $f(\mathbf{x})$ with respect to x_i:

First derivative:

$$\frac{\partial f(\mathbf{x})}{\partial x_i} = \frac{\exp(x_i)}{\sum_{j=1}^{n} \exp(x_j)}$$

Second derivative:

$$\frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} = \frac{\exp(x_i)\exp(x_j)}{\left(\sum_{k=1}^{n} \exp(x_k)\right)^2} - \delta_{ij}\frac{\exp(x_i)}{\sum_{k=1}^{n} \exp(x_k)}$$

where δ_{ij} is the Kronecker delta, which is equal to 1 if $i = j$ and 0 otherwise.

Let's examine the Hessian matrix, $H \in \mathbb{R}^{n\times n}$, where $H_{ij} = \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j}$, which is the second derivative of the log-sum-exp function. We want to show that H is positive semi-definite.

Let $\mathbf{v} \in \mathbb{R}^n$ be an arbitrary non-zero vector. We have to show that $\mathbf{v}^\top H\mathbf{v} \geq 0$ for all $\mathbf{v}$.

$$\mathbf{v}^\top H\mathbf{v} = \sum_{i=1}^{n}\sum_{j=1}^{n} v_i \left(\frac{\exp(x_i)\exp(x_j)}{\left(\sum_{k=1}^{n} \exp(x_k)\right)^2} - \delta_{ij}\frac{\exp(x_i)}{\sum_{k=1}^{n} \exp(x_k)}\right) v_j$$

Let's define:

$$A = \sum_{i=1}^{n} v_i \frac{\exp(x_i)}{\sum_{k=1}^{n} \exp(x_k)}$$

Thus, we can rewrite the expression as:
$\mathbf{v}^\top H\mathbf{v} = A^2 - B$
where:

$$B = \sum_{i=1}^{n} v_i^2 \frac{\exp(x_i)}{\sum_{k=1}^{n} \exp(x_k)}$$

Using the Cauchy-Schwarz inequality, we have:

$$\left(\sum_{i=1}^{n} v_i \frac{\exp(x_i)}{\sum_{k=1}^{n} \exp(x_k)}\right)^2 \leq \left(\sum_{i=1}^{n} v_i^2\right)\left(\sum_{i=1}^{n}\left(\frac{\exp(x_i)}{\sum_{k=1}^{n} \exp(x_k)}\right)^2\right)$$

We can divide it on both sides of the inequality by $\sum_{i=1}^{n} v_i^2$, obtaining:

$$\sum_{i=1}^{n} \frac{\left(\frac{\exp(x_i)}{\sum_{k=1}^{n} \exp(x_k)}\right)^2}{\sum_{i=1}^{n} v_i} \leq \sum_{i=1}^{n}\left(\frac{\exp(x_i)}{\sum_{k=1}^{n} \exp(x_k)}\right)^2$$

We can multiply $\sum_{i=1}^{n} v_i^2 > 0$ on both sides of the inequality, obtaining:

$$A^2 \leq B$$

Now, we have shown that $A^2 \leq B$, which means that $\mathbf{v}^\top H\mathbf{v} = A^2 - B \geq 0$. Therefore, the Hessian matrix H is positive semi-definite, and the log-sum-exp function $f(\mathbf{x})$ is convex. □

the objective of logistic regression for binary classification convexity

Proof. To prove that the objective function of logistic regression for binary classification $f(w) = \frac{1}{n}\sum_{i=1}^{n}\log(1+\exp(-y_i w^\top x_i)) + \frac{\lambda}{2}\|w\|_2^2$ is convex, we will first prove the convexity of $h(s) = \log(1+\exp(s))$.

First let's compute the first and second derivatives of $h(s)$.
First derivative:
$$\frac{dh(s)}{ds} = \frac{1}{1+\exp(s)} \cdot \exp(s)$$

Second derivative:
$$\frac{d^2h(s)}{ds^2} = \frac{d}{ds}\left(\frac{\exp(s)}{1+\exp(s)}\right) = \frac{\exp(s)(1+\exp(s)) - \exp^2(s)}{(1+\exp(s))^2} = \frac{\exp(s)}{(1+\exp(s))^2}$$

Notice that the denominator is $(1+\exp(s))^2$ is always positive for all $s \in \mathbb{R}$, so we only need to analyze the numerator $\exp(s)$. Since the exponential function $\exp(s)$ is always positive, so the second derivative is non-negative for all real values of s, which means that the function $h(s) = \log(1+\exp(s))$ is convex.
By composition with affine function rule that preserves convexity,
$$\log(1+\exp(-y_i w^\top x_i))$$
is convex. So
$$\frac{1}{n}\sum_{i=1}^{n} \log(1+\exp(-y_i w^\top x_i))$$
is convex based on the non-negative weighted sum rule and non-negative multiple rules that preserve convexity.
$\|w\|_2^2$ is convex from the slide (least-squares objective), and $\frac{\lambda}{2}\|w\|_2^2$ is convex based on non-negative multiple rule that preserves convexity.
So
$$\frac{1}{n}\sum_{i=1}^{n} \log(1+\exp(-y_i w^\top x_i)) + \frac{\lambda}{2}\|w\|_2^2$$
is convex based on non-negative weighted sum rule that preserves convexity.
Therefore, the objective function of logistic regression for binary classification $f(w)$ is convex.

□

the objective of support vector machine convexity

Proof. To prove that the objective function $f(w) = \frac{1}{n}\sum_{i=1}^{n} \max(0, 1-y_i w^\top x_i) + \frac{\lambda}{2}\|w\|_2^2$ of the support vector machine is convex, we will first establish the convexity of $q_i(s) = \max(0, 1-y_i s^\top x_i)$.
Let $g(s) = y_i s^\top x_i$ and $h(s) = 0$. Both $g(s)$ and $h(s)$ are affine functions with respect to s, so they are convex (affine functions are both convex and concave).

Now $q_i(s) = \max(g(s), h(s))$. Based on the pointwise maximum rule that preserves convexity, the pointwise maximum of convex functions is convex, thus $q_i(s)$ is convex.
We just showed $\max(0, 1 - y_i w^\intercal x_i)$ is convex, and

$$\frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - y_i w^\intercal x_i)$$

is convex based on the non-negative weighted sum rule and non-negative multiple rules that preserve convexity.
In a previous proof, we have shown $\frac{\lambda}{2}\|w\|_2^2$ is convex. So

$$\frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - y_i w^\intercal x_i) + \frac{\lambda}{2}\|w\|_2^2$$

is convex based on the non-negative weighted sum rule that preserves convexity.
Therefore, the objective function of the support vector machine $f(w)$ is convex. □

2 the constrained version of Ridge Regression

Does strong duality hold?

Yes.

Proof. To analyze whether strong duality holds for the constrained ridge regression problem, we can check if the problem satisfies Slater's constraint qualification (SCQ). The ridge regression problem is a convex optimization problem, so if it satisfies the SCQ, then strong duality holds.
The constrained ridge regression problem can be written as:

$$\min_{w\in\mathbb{R}^d} \frac{1}{2}\|\Phi w - y\|_2^2 \quad \text{s.t.} \quad \|w\|_2^2 \le s$$

Here, the objective function is convex since it is a quadratic function with positive coefficients, and the constraint is also convex as it represents a Euclidean ball. Slater's constraint qualification requires the existence of a strictly feasible point, i.e., a point that strictly satisfies the inequality constraint. In this case, we can choose $w = 0$, which satisfies the constraint as $\|0\|_2^2 = 0 < s$, assuming $s > 0$. Thus, the SCQ is satisfied, and strong duality holds.
Now derive the KKT conditions for the optimal solution w^*. The Lagrangian of the problem is given by:

$$\mathcal{L}(w, v) = \frac{1}{2}\|\Phi w - y\|_2^2 + v\left(\frac{1}{2}\|w\|_2^2 - s\right)$$

where $v \geq 0$ is the Lagrange multiplier. The KKT conditions consist of the following:

1. The gradient of Lagrangian with respect to $\Phi^\top(\Phi w - y)$ vanishes:

$$\nabla_w \mathscr{L}(w, v) = \Phi^\top(\Phi w - y) + vw = 0$$

2. Primal constraints:

$$\|w\|_2^2 \leq s$$

3. Dual constraints:

$$v \geq 0$$

4. Complementary slackness:

$$v\left(\frac{1}{2}\|w\|_2^2 - s\right) = 0$$

These KKT conditions characterize the optimal solution w^* of the constrained ridge regression problem. Solving these conditions can help find the optimal w^* and the corresponding Lagrange multiplier v. □

Does a close-formed solution exist?

For the constrained ridge regression problem, there isn't a closed-form solution like in the unconstrained ridge regression case. However, we can use an algorithm to compute the optimal solution.

1. Initialize point and step size
2. Calculate the gradient of the objective function at the current point.
3. Update the point using the gradient.
4. Project the updated point onto the constraint set.
5. If the algorithm has converged, stop. Otherwise, continue for another iteration and go back to step 2.

3 The equivalence between Max Entropy Model and the Logistic Regression

We want to show that the Maximum Entropy Model is equivalent to the multi-class logistic regression model.

Proof. To show that the Maximum Entropy Model with feature function $f_j(x_i) = [x_i]_j$ is equivalent to the multi-class logistic regression model without regularization.
The optimization problem for the Maximum Entropy Model is

$$\begin{aligned}
&\max_{p(y|x_i)} -\sum_{i=1}^{n}\sum_{y=1}^{K} p(y|x_i)\ln p(y|x_i) \\
&\text{s.t.} \sum_{y=1}^{K} p(y|x_i) = 1 \\
&\frac{1}{n}\sum_{i=1}^{n}\delta(y,y_i)[x_i]_j = \frac{1}{n}\sum_{i=1}^{n} p(y|x_i)[x_i]_j, \quad j=1,\dots,d, \quad y=1,\dots,K
\end{aligned}$$

where $\delta(y, y_i)$ is equal to 1 if $y_i = y$, and 0 otherwise.
We use the Lagrangian dual theory to solve the constrained optimization problem:

$$\begin{aligned}
\mathscr{L}(p,\lambda,\mu) = &-\sum_{i=1}^{n}\sum_{y=1}^{K} p(y|x_i)\ln p(y|x_i) + \sum_{i=1}^{n}\lambda_i\left(\sum_{y=1}^{K} p(y|x_i) - 1\right) \\
&+ \sum_{y=1}^{K}\sum_{j=1}^{d}\mu_{j,y}\left(\frac{1}{n}\sum_{i=1}^{n}\delta(y,y_i)[x_i]_j - \frac{1}{n}\sum_{i=1}^{n} p(y|x_i)[x_i]_j\right)
\end{aligned}$$

Taking the derivative with respect to $p(y|x_i)$ and setting it to zero, we get:

$$\frac{\partial\mathscr{L}}{\partial p(y|x_i)} = -\ln p(y|x_i) - 1 + \lambda_i + \frac{\mu_{j,y}[x_i]_j}{n} = 0$$

Then solving for $p(y|x_i)$:

$$p(y|x_i) = \exp\left(\lambda_i + \frac{\mu_{j,y}[x_i]_j}{n} - 1\right)$$

To satisfy the constraint $\sum_{y=1}^{K} p(y|x_i) = 1$, we can normalize the probabilities:

$$p(y|x_i) = \frac{\exp\left(\lambda_i + \frac{\mu_{j,y}[x_i]_j}{n} - 1\right)}{\sum_{y'=1}^{K} \exp\left(\lambda_i + \frac{\mu_{j,y'}[x_i]_j}{n} - 1\right)}$$

Comparing this expression of the multi-class logistic regression model:

$$p(y|x_i) = \frac{\exp(w_y^\top x_i)}{\sum_{y'=1}^{K} \exp(w_{y'}^\top x_i)}$$

We can see that both expressions have the same form, where the correspondence between the parameters is:

$$\frac{\mu_{j,y}[x_i]_j}{n} = w_y^\top x_i$$

Therefore, we have shown that the Maximum Entropy Model with the given feature function is equivalent to the multi-class logistic regression model without regularization. □