INFR10086 Machine Learning (MLG) Semester 2, 2025/6

Coursework

Instructions

- Due date: 9 March, Monday, at 12pm
- The submission is through Gradescope `https://www.gradescope.com/courses/1224477/assignments/7455000`.
- It's best to typeset your answers, but it is fine to submit hand-written answers.

Questions

1. In this question, we will look at the relationship between the hinge loss and support vector machines (SVM).

   Given a data set $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$, soft-margin SVM is defined as the following optimization problem

$$\begin{aligned} \min \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & 1 - y_i\mathbf{w}^\top\mathbf{x}_i \le \xi_i \text{ for } i = 1, \dots, n \\ & \xi_i \ge 0 \text{ for } i = 1, \dots, n \end{aligned} \tag{1}$$

   where C is a hyperparameter.[1] The hinge loss for a single sample (x, y) is defined as $\max(0, 1 - y\mathbf{w}^\top\mathbf{x})$. Minimizing the hinge loss on the entire data set becomes

$$\frac{\lambda}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n}\max(0, 1 - y_i\mathbf{w}^\top\mathbf{x}_i), \tag{2}$$

   where λ is another hyperparameter.

   (a) Show that $\xi_i = \max(0, 1 - y\mathbf{w}^\top\mathbf{x}_i)$. In other words, the slack variable ξ is the hinge loss for the data point $(\mathbf{x}_i, y_i)$, and thus soft-margin SVM is equivalent to minimizing the hinge loss.

   [5 marks]

[1] A hyperparameter is not something we optimize as part of the optimization problem, but rather a constant of our choice. In practice, we typically try a few a hyperparameters and choose the best.

(b) Find the Lagrangian L of soft-margin SVM and solve $\nabla_{\mathbf{w}} L = 0$. In particular, if α_i is the Lagrange multiplier for $1 - y_i \mathbf{w}^\top \mathbf{x}_i \leq \xi_i$, then

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i. \tag{3}$$

The optimal $\mathbf{w}$ is a linear combination of data points from the data set.[2]

[5 marks]

(c) Because the optimal solution needs to satisfy complementary slackness, show that when

$$1 - y_i \mathbf{w}^\top \mathbf{x}_i < 0 \tag{4}$$

then $\alpha_i = 0$.

[5 marks]

(d) Use the above and conclude that any point $(\mathbf{x}_i, y_i)$ such that

$$1 - y_i \mathbf{w}^\top \mathbf{x}_i < 0 \tag{5}$$

are not part of the optimal $\mathbf{w}$. In particular, the optimal solution does not change if some of these points are removed from the data set.

[5 marks]

2. In this question, we will prove a convergence result for gradient descent.

(a) A function f is L-Lipschitz if

$$f(\mathbf{x}) - f(\mathbf{y}) \leq L\|\mathbf{x} - \mathbf{y}\| \tag{6}$$

for all $\mathbf{x}$ and $\mathbf{y}$. A function f is L-smooth if its gradient is L-Lipschitz. Show that when a function is both convex and L-smooth, then

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + L\|\mathbf{x} - \mathbf{y}\|^2. \tag{7}$$

You will need to use the Cauchy-Shwarz inequality, which states that $\mathbf{x}^\top \mathbf{y} \leq \|\mathbf{x}\|\|\mathbf{y}\|$ for all $\mathbf{x}$ and $\mathbf{y}$.

[5 marks]

(b) Consider doing gradient descent

$$\mathbf{x}_t = \mathbf{x}_{t-1} - \eta_t \nabla f(\mathbf{x}_{t-1}) \tag{8}$$

for $t = 1, \dots, k$, on a convex and L-smooth function f.

i. Start with equation (7), use the definition of gradient descent, and show that

$$f(\mathbf{x}_{t-1}) - f(\mathbf{x}_t) \geq \frac{1}{4L}\|\nabla f(\mathbf{x}_{t-1})\|^2, \tag{9}$$

when $\eta_t = \frac{1}{2L}$. Note that the norm is always nonnegative, and our objective, in this case, always decreases. This result is commonly known as the descent lemma.

[2 marks]

[2] This result is itself important and has a name called the representer theorem.

ii. Expand $\|\mathbf{x}_t - \mathbf{x}^*\|^2$ and use the definition of gradient descent to show

$$\nabla f(\mathbf{x}_{t-1})^\top (\mathbf{x}_{t-1} - \mathbf{x}^*) = L(\|\mathbf{x}_{t-1} - \mathbf{x}^*\|^2 - \|\mathbf{x}_t - \mathbf{x}^*\|^2) + \frac{1}{4L}\|\nabla f(\mathbf{x}_{t-1})\|^2, \quad (10)$$

where $\mathbf{x}^*$ is the optimal solution.

[2 marks]

iii. Use the descent lemma and show

$$\nabla f(\mathbf{x}_{t-1})^\top (\mathbf{x}_{t-1} - \mathbf{x}^*) \leq L(\|\mathbf{x}_{t-1} - \mathbf{x}^*\|^2 - \|\mathbf{x}_t - \mathbf{x}^*\|^2) + f(\mathbf{x}_{t-1}) - f(\mathbf{x}_t). \quad (11)$$

[1 mark]

iv. Use convexity and equation (11) to show

$$f(\mathbf{x}_{t-1}) - f(\mathbf{x}^*) \leq L(\|\mathbf{x}_{t-1} - \mathbf{x}^*\|^2 - \|\mathbf{x}_t - \mathbf{x}^*\|^2) + f(\mathbf{x}_{t-1}) - f(\mathbf{x}_t). \quad (12)$$

In particular,

$$f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq L(\|\mathbf{x}_{t-1} - \mathbf{x}^*\|^2 - \|\mathbf{x}_t - \mathbf{x}^*\|^2). \quad (13)$$

[1 mark]

v. Take the sum of t from 1 to k on both sides of equation (13) (and divide it by k) to get

$$\frac{1}{k}\sum_{t=1}^{k} f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \frac{L}{k}(\|\mathbf{x}_0 - \mathbf{x}^*\|^2 - \|\mathbf{x}_k - \mathbf{x}^*\|^2) \leq \frac{L}{k}\|\mathbf{x}_0 - \mathbf{x}^*\|^2. \quad (14)$$

[1 mark]

vi. Because of the descent lemma, show that

$$f(\mathbf{x}_k) \leq \frac{1}{k}\sum_{t=1}^{k} f(\mathbf{x}_t). \quad (15)$$

[1 mark]

vii. Finally, putting everything together, we have

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{L}{k}\|\mathbf{x}_0 - \mathbf{x}^*\|^2, \quad (16)$$

after running k steps of gradient descent on a convex and L-smooth function f.

[1 mark]

viii. Given the above, what is the convergence rate of gradient descent on a convex and L-smooth function?

[1 mark]

3. In programming, we have various programming constructs to work with, such as if statements and for loops. In this question, we will look at learning the following if statement.

if $\mathbf{x}$ has property ζ **then**
  return $f(\mathbf{x})$
else
  return $g(\mathbf{x})$
end if

We can let $z = +1$ when $\mathbf{x}$ has property ζ, and let $z = -1$ otherwise. The if statement can then be rewritten into

$$\text{cond}(\mathbf{x}, f, g) = \begin{cases} f(\mathbf{x}) & \text{if } p(z = +1|\mathbf{x}) \geq 0.5 \\ g(\mathbf{x}) & \text{if } p(z = +1|\mathbf{x}) < 0.5 \end{cases} \tag{17}$$

Suppose the result of cond is send immediately to a loss function L, i.e., computing $L(\text{cond}(x, f, g), y)$ for a labeled sample (x, y).

Because we do not know ahead of time whether z is going to be $+1$ or -1, we can only measure the loss in expectation

$$\mathbb{E}_z[L(\text{cond}(\mathbf{x}, f, g), y)] = p(z = +1|\mathbf{x})L(f(\mathbf{x}), y) + p(z = -1|\mathbf{x})L(g(\mathbf{x}), y). \tag{18}$$

(a) Suppose we parameterize $p(z|\mathbf{x})$ as

$$p(z = +1|\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^\top \mathbf{x})}. \tag{19}$$

Show that

$$\nabla_{\mathbf{w}} \mathbb{E}_z[L(\text{cond}(\mathbf{x}, f, g), y)] = p(z = +1|\mathbf{x})p(z = -1|\mathbf{x})(L(f(\mathbf{x}), y) - L(g(\mathbf{x}), y))\mathbf{x}. \tag{20}$$

In particular, learning $\mathbf{w}$ only requires computing $f(\mathbf{x})$ and $g(\mathbf{x})$ but does not require backpropagation through f and g.

[5 marks]

(b) If f and g happen to have the same return type, we have

$$\mathbb{E}_z[\text{cond}(\mathbf{x}, f, g)] = p(z = +1|\mathbf{x})f(\mathbf{x}) + p(z = -1|\mathbf{x})g(\mathbf{x}). \tag{21}$$

When L is convex in the first argument, show that

$$L(\mathbb{E}_z[\text{cond}(\mathbf{x}, f, g)], y) \leq \mathbb{E}_z[L(\text{cond}(\mathbf{x}, f, g), y)]. \tag{22}$$

[5 marks]

(c) Discuss the pros and cons of measuring $L(\mathbb{E}_z[\text{cond}(\mathbf{x}, f, g)], y)$ or $\mathbb{E}_z[L(\text{cond}(\mathbf{x}, f, g), y)]$, in particular, when are they equal and which one we should use when they are equal.

[5 marks]