Exercises Set: Binary Classification

DSBA Mathematics Refresher 2024

Abstract

As this is the last session, there will be no compulsory questions this time.

1 Lagrangian multiplier technique

1.1 Unconstrained optimization

Let $f(x, y) = 2x^2 - 12x + 4y^2 + 8y + 20$.
Find $(x^*, y^*) \in \mathbb{R}^2$ such that f reaches its minimum (i.e. $f(x^*, y^*) \leq f(x, y) \quad \forall (x, y) \in \mathbb{R}^2$).

1.2 (Equality) Constrained optimization

Let $f(x, y) = 2x^2 - 12x + 4y^2 + 8y + 20$.
Suppose further that we want $3x + 5y = 2$.
Find $(x^*, y^*) \in \mathbb{R}^2$ such that $3x^* + 5y^* = 2$ and f reaches its minimum (i.e. $f(x^*, y^*) \leq f(x, y) \quad \forall (x, y) \in \mathbb{R}^2,\ 3x + 5y = 2$).

1.3 Lagrange multiplier

Let $f(x, y) = 2x^2 - 12x + 4y^2 + 8y + 20$.
Suppose further that we want $3x + 5y = 2$.
Let $\mathcal{L}(x, y, \lambda) = f(x, y) - \lambda(3x + 5y - 2)$.
Find the point where $\nabla \cdot \mathcal{L} = 0$

1.4 (Inequality) Constrained optimization

Let $f(x) = x^2 - 2x$.
Suppose further that we want $3x \leq 2$.
Find $x^* \in \mathbb{R}$ such that $3x^* \leq 2$ and f reaches its minimum (i.e. $f(x^*) \leq f(x) \quad \forall x \in \mathbb{R},\ 3x \leq 2$).

Let $f(x) = x^2 + 2x$.
Suppose further that we want $3x \leq 2$.
Find $x^* \in \mathbb{R}$ such that $3x^* \leq 2$ and f reaches its minimum (i.e. $f(x^*) \leq f(x) \quad \forall x \in \mathbb{R},\ 3x \leq 2$).

1.5 Lagrange multiplier

Let $f(x) = x^2 - 2x$.
Suppose further that we want $3x \leq 2$.
Let $\mathcal{L}(x, \lambda) = f(x) - \lambda(-3x + 2 - s^2)$.
Find the point where $\nabla \cdot \mathcal{L} = 0$ and $\lambda \geq 0$.

Let $f(x) = x^2 + 2x$.
Suppose further that we want $3x \leq 2$.
Let $\mathcal{L}(x, \lambda) = f(x) - \lambda(-3x + 2 - s^2)$.
Find the point where $\nabla \cdot \mathcal{L} = 0$ and $\lambda \geq 0$.

2 Support Vector Machines

2.1 Theory

Define a line in $\mathbb{R}^2$ with parameters $\vec{\mathbf{w}} = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$ and b [1] defined by $\vec{\mathbf{w}}.\vec{\mathbf{x}} = b$ (or $\vec{\mathbf{w}}.\vec{\mathbf{x}} - b = 0$) for $\vec{\mathbf{x}} = \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2$. This line cut the plane in 2 regions:

- $\mathcal{R}_{-1}$: $\vec{\mathbf{x}} \in \mathbb{R}^2$ such that $\vec{\mathbf{w}}.\vec{\mathbf{x}} - b < 0$
- $\mathcal{R}_{+1}$: $\vec{\mathbf{x}} \in \mathbb{R}^2$ such that $\vec{\mathbf{w}}.\vec{\mathbf{x}} - b > 0$

The goal is to find $\vec{\mathbf{w}}$ and b such that all points of the first class are in the first region, and all points of the second class are in the second region.

Let our data be $\{(\vec{\mathbf{x_k}}, y_k)\}_{k=1}^{N}$ where $\vec{\mathbf{x_k}}$ is the coordinate of the point in $\mathbb{R}^2$ and y_k is the label ($+1$ or -1).

The best line is not only separating the data in two sets, but also maximizing the distance between the line and the points.
TODO: Calculate the distance between the line and a point $\vec{\mathbf{x}} \in \mathbb{R}^2$.

Let $h(\vec{\mathbf{x}}) = \vec{\mathbf{w}}^T \cdot \vec{\mathbf{x}} - b$. We will predict first class -1 if $h(\vec{\mathbf{x}}) < 0$ (as this means $\vec{\mathbf{x}} \in \mathcal{R}_{-1}$); and second class $+1$ if $h(\vec{\mathbf{x}}) > 0$ (as this means $\vec{\mathbf{x}} \in \mathcal{R}_{+1}$). To have each item of data classified correctly, we need:

- if $y_k = +1$ then $h(\vec{\mathbf{x_k}}) > 0$
- if $y_k = -1$ then $h(\vec{\mathbf{x_k}}) < 0$

TODO: Derive a condition that ensures correct classification if true for all k.

Assuming the data is linearly separable, we now define 3 regions:

- $\mathcal{R}_{-1}$: $\vec{\mathbf{x}} \in \mathbb{R}^2$ such that $\vec{\mathbf{w}}.\vec{\mathbf{x}} - b < -1$
- $\mathcal{R}_0$: $\vec{\mathbf{x}} \in \mathbb{R}^2$ such that $\mid \vec{\mathbf{w}}.\vec{\mathbf{x}} - b \mid < 1$ *(the "margin band")*
- $\mathcal{R}_{+1}$: $\vec{\mathbf{x}} \in \mathbb{R}^2$ such that $\vec{\mathbf{w}}.\vec{\mathbf{x}} - b > 1$

TODO: Calculate the margin band width.
TODO: Formulate the constrained optimization problem to find the best (i.e. maximizing the margin) parameters $\vec{\mathbf{w}}$ and b.

2.2 Practice

The training dataset consists of the following data points:

Positive class ("+1"):

[1] Here, $\vec{\mathbf{w}} = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$, $\vec{\mathbf{x}} = \begin{pmatrix} x \\ y \end{pmatrix}$, and b is a scalar. We use this notation because it makes it very easy to extend the technique to higher dimensions. For example, you can just add one component to use a plane to separate points in $\mathbb{R}^3$

- (2, 2)
- (1, 1)

Negative class ("-1"):

- (0, 1)
- (1, 0)

1. Plot the points in a 2D graph.
2. Using the graph, find the optimal $\vec{\mathbf{w}}$.
3. Calculate the optimal value of b to separate the data points while maximizing the margin.
4. Determine the equation of the optimal hyperplane in the form $\vec{\mathbf{w}} \cdot \vec{\mathbf{x}} + b = 0$.
5. Identify the support vectors in the dataset.
6. Calculate the margin, which is the perpendicular distance from the hyperplane to the nearest support vector.
7. Classify a new data point, $(3, 2)$, based on the learned SVM model.

Remarks:
Here, we have made educated guesses to find $\vec{\mathbf{w}}$ and b; in practice, we need to solve the constrained optimization problem stated above.
This can be done using Lagrange multiplier, the details are outside the scope of this course.
In practice, we let computers solve the problem for us.

2.3 Kernel

Suppose we have the following data: Positive class ("+1"):

- (0, 1)
- (0, -1)
- (1, 0)
- (-1, 0)

Negative class ("-1"):

- (0, 2)
- (0, -2)
- (2, 0)

- (-2, 0)

Plot the data points.
Observe that there is no line separating the two classes perfectly.

Let $\phi(x, y) = (x^2 + y^2, x^2 - y^2)$. Compute $\phi(x, y)$ for every point, and plot them on a new graph.
Observe that the two classes should now be separable by a line [2].

3 Logistic Regression

You are working on a binary classification problem where you are using logistic regression to predict whether a student will pass (1) or fail (0) an exam based on the number of hours they have studied. You have collected the data in the provided table.

Hours Studied (X)	Pass (Y)
0	0 (no)
4	0 (no)
3	1 (yes)
8	1 (yes)

You want to fit a logistic regression model to this data and find the best-fitting sigmoid function, which is represented as:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

Where $P(Y = 1|X)$ is the probability of passing the exam given the hours studied.

1. Calculate the cost function (log loss) for the given data and the predicted probabilities. The log loss for a single data point is given as:

$$\text{Log Loss} = -\left(Y \cdot \log(P(Y = 1|X)) + (1 - Y) \cdot \log(1 - P(Y = 1|X))\right)$$

2. Calculate the total cost (log loss) summing for all data points.

3. Use gradient descent or any other optimization method to find the values of β_0 and β_1 that minimize the total cost; you may use a computer to perform gradient descent.

4. Calculate the probability $P(Y = 1|X = 10)$ using the logistic regression model[3].

[2] The "kernel trick" allows us to perform an SVM on data transformed by a non-linear function; the details are outside the scope of this course.

[3] That is, the estimated probability of passing given the fact that the student have been studying for ten hours.

5. Calculate the probability $P(Y = 1|X = 1)$ using the logistic regression model[4].

6. Calculate the probability $P(Y = 1|X = 0)$ using the logistic regression model[5].

[4]That is, the estimated probability of passing given the fact that the student have been studying for one hour.

[5]That is, the estimated probability of passing given the fact that the student have not studied at all.