Lecture 5: Introduction to Machine Learning

Yonghao Lee

December 2, 2025

1 Linear Hypothesis Classes

A linear hypothesis takes the form of a linear decision rule. In a binary classification setting, we predict positive if:

$$\mathbf{w} \cdot \mathbf{x} + b > 0 \tag{1}$$

Visually, this represents a hyperplane separating the data space.

1.1 Convexity

To find the best $\mathbf{w}$ and b, we optimize a loss function. We prefer **convex functions** (shaped like a bowl) because a local minimum in a convex function is guaranteed to be a global minimum.

2 Optimization: Gradient Descent

Since we cannot always find the optimal weights $\mathbf{w}$ by simple guessing, we use an iterative algorithm called Gradient Descent.

2.1 The Algorithm

We start with an initial $\mathbf{w}_{init}$ and iteratively move in the direction opposite to the gradient (the direction of steepest descent).

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \alpha \nabla_{\mathbf{w}} L(\mathbf{w}_t) \tag{2}$$

Where:

- $\nabla_{\mathbf{w}} L$: The gradient of the loss function.
- α: The step size (learning rate).

2.2 Step Size Intuition

The step taken is proportional to the gradient magnitude.

- **Steep slope (large gradient):** We take a large step.
- **Flat slope (small gradient):** We take a small step to settle into the minimum.

2.3 Stochastic Gradient Descent (SGD)

Computing the gradient over the entire training set (N_{train}) is computationally expensive.

- **SGD:** Approximates the gradient using a *single* sample per step.
- **Mini-batch GD:** Approximates the gradient using a small batch of samples.

3 Linear Regression

The task of regression is to predict a continuous scalar output y given an input $\mathbf{x}$.

3.1 Squared Loss

We minimize the Empirical Risk, defined as the Mean Squared Error (MSE):

$$L(S_{train}, \mathbf{w}) = \frac{1}{2N_{train}} \sum_{(x,y)\in S_{train}} (y - \mathbf{w}\cdot\mathbf{x})^2 \tag{3}$$

The gradient of this loss function with respect to $\mathbf{w}$ is:

$$\nabla_{\mathbf{w}} L = \frac{1}{N_{train}} \sum_{(x,y)\in S_{train}} \mathbf{x}(\mathbf{w}\cdot\mathbf{x} - y) \tag{4}$$

4 L2 Regularization (Ridge Regression)

To prevent overfitting and control model complexity, we introduce L2 Regularization. We penalize the magnitude of the weights.

4.1 Regularized Loss Function

The new loss function includes a penalty term weighted by hyperparameter α:

$$L_{total}(\mathbf{w}) = \underbrace{\frac{1}{2N_{train}} \sum (y - \mathbf{w}\cdot\mathbf{x})^2}_{\text{Data Term}} + \underbrace{\frac{\alpha}{2}\|\mathbf{w}\|^2}_{\text{Regularization}} \tag{5}$$

The gradient becomes:

$$\nabla L = \frac{1}{N_{train}} \sum \mathbf{x}(\mathbf{w}\cdot\mathbf{x} - y) + \alpha\mathbf{w} \tag{6}$$

4.2 The Analytical Solution

Unlike standard gradient descent, Ridge Regression allows for a **closed-form solution** by setting the gradient to zero and solving using matrix algebra.

Let X be the data matrix and $\mathbf{y}$ be the label vector.

$$\nabla L = 0 \tag{7}$$

$$X^T(X\mathbf{w} - \mathbf{y}) + \alpha I\mathbf{w} = 0 \tag{8}$$

$$X^T X\mathbf{w} + \alpha I\mathbf{w} = X^T\mathbf{y} \tag{9}$$

$$(X^T X + \alpha I)\mathbf{w} = X^T\mathbf{y} \tag{10}$$

The optimal weights $\mathbf{w}^*$ are:

$$\mathbf{w}^* = (X^T X + \alpha I)^{-1} X^T\mathbf{y} \tag{11}$$

Note: The term αI ensures the matrix is invertible (non-singular).

5 Classification

5.1 Logistic Regression

Instead of predicting a raw scalar, we predict the probability $P(y = 1|\mathbf{x})$. We use the **Sigmoid Function** to squash the score into $[0, 1]$:

$$\sigma(s) = \frac{1}{1 + e^{-s}} \tag{12}$$

The loss function is the **Negative Log Likelihood** (Log Loss):

$$\ell(\mathbf{w}, x, y) = -\log(p_{\mathbf{w}}(x)) \quad \text{for the true class} \tag{13}$$

5.1.1 Multi-Class Extension

For K classes, we maintain K weight vectors and use the **Softmax** function:

$$\text{Soft}(k|\mathbf{z}) = \frac{e^{z^k}}{\sum_{j=1}^{K} e^{z^j}} \tag{14}$$

5.2 Ridge Classification

A simpler alternative to Logistic Regression. We treat the binary labels (e.g., $y \in \{+1, -1\}$) as real numbers and apply Ridge Regression.

- **Method:** Minimize squared error between raw prediction and categorical label.
- **Advantage:** Computationally fast because it uses the analytical solution derived in Section 4.
- **Performance:** Often yields results comparable to Logistic Regression despite treating classification as a regression problem.