Study Guide: Binary Classification and Logistic Regression

Student Name

October 9, 2024

1 Binary Classification (Section 1)

In binary classification, the target label y takes values from $\{-1, +1\}$, where $y = +1$ represents the positive class and $y = -1$ represents the negative class. The task is to classify a data point $x \in \mathbb{R}^n$ based on a hypothesis $h_\theta(x) = \theta^\top x$, where $\theta \in \mathbb{R}^n$ is the parameter vector.

1.1 Classification Rule

The predicted class is determined by the sign of the linear combination of features:

$$\text{sign}(h_\theta(x)) = \text{sign}(\theta^\top x),$$

where

$$\text{sign}(t) = \begin{cases} 1 & \text{if } t > 0, \\ 0 & \text{if } t = 0, \\ -1 & \text{if } t < 0. \end{cases}$$

A hypothesis h_θ classifies an example (x, y) correctly if:

$$y\theta^\top x > 0.$$

The quantity $y\theta^\top x$ is known as the **margin**, and it is used as a measure of confidence in the classification. A large positive margin indicates a confident correct classification.

2 Loss Functions

In binary classification, the choice of loss function is critical. The goal is to minimize a loss function ϕ that penalizes incorrect classifications. The empirical risk minimized over the training set $\{(x^{(i)}, y^{(i)})\}_{i=1}^m$ is:

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m \phi(y^{(i)}\theta^\top x^{(i)}).$$

We desire a loss function $\phi(z)$ such that:

$$\phi(z) \to 0 \text{ as } z \to \infty, \quad \phi(z) \to \infty \text{ as } z \to -\infty.$$

This ensures that large positive margins (correct classifications) incur low loss, while negative margins (misclassifications) incur high loss.

2.1 Zero-One Loss

The **zero-one loss** is defined as:

$$\phi_{\text{zo}}(z) = \begin{cases} 1 & \text{if } z \leq 0, \\ 0 & \text{if } z > 0. \end{cases}$$

However, this loss is non-convex and discontinuous, making it hard to minimize in practice.

2.2 Common Loss Functions

Three commonly used loss functions in machine learning are:

- **Logistic Loss:**
$$\phi_{\text{logistic}}(z) = \log(1 + e^{-z}).$$
- **Hinge Loss:**
$$\phi_{\text{hinge}}(z) = \max(0, 1 - z).$$
- **Exponential Loss:**
$$\phi_{\exp}(z) = e^{-z}.$$

Each of these loss functions is convex and ensures that the loss decreases as the margin increases.

3 Logistic Regression (Section 2)

Logistic regression is a classification algorithm that minimizes the logistic loss function. The empirical risk minimized by logistic regression is:

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \log\left(1 + e^{-y^{(i)}\theta^\top x^{(i)}}\right).$$

The goal is to find the parameter vector θ that minimizes this loss.

3.1 Probabilistic Interpretation

Logistic regression can be interpreted probabilistically by introducing the **sigmoid function**:

$$g(z) = \frac{1}{1 + e^{-z}}.$$

The probability that $y = 1$ given x is modeled as:

$$p(y = 1|x; \theta) = g(\theta^\top x).$$

The log-likelihood of the training data is:

$$\ell(\theta) = \sum_{i=1}^{m} \log g(y^{(i)}\theta^\top x^{(i)}),$$

which is equivalent to minimizing the logistic loss.

3.2 Gradient Descent for Logistic Regression

The gradient of the logistic loss for a single example (x, y) is:

$$\nabla_\theta \phi_{\text{logistic}}(yx^\top \theta) = -g(-yx^\top \theta) yx.$$

The **stochastic gradient descent** update rule for logistic regression is:

$$\theta^{(t+1)} = \theta^{(t)} + \alpha_t g(-y^{(i)} x^{(i)} \theta^{(t)}) y^{(i)} x^{(i)},$$

where α_t is the learning rate at iteration t.

4 Summary

- Binary classification uses a linear classifier $h_\theta(x) = \theta^\top x$, with predictions based on the sign of the linear combination of features.
- The margin $y\theta^\top x$ measures confidence in the classification, with a large positive margin indicating a confident correct classification.
- Several loss functions can be used to train binary classifiers, including logistic loss, hinge loss, and exponential loss.
- Logistic regression minimizes the logistic loss using gradient-based methods, with the sigmoid function providing a probabilistic interpretation.