CS 412 — Introduction to Machine Learning (UIC) February 04, 2025

Lecture 4

Lecturer: Aadirupa Saha *Scribe(s): Sujay Dahagam*

Overview

Previous Lecture (Lecture 3) Topics:

1. **Revisiting Linear Regression (for $d = 1$):** The basic linear model $y = wx + b$ and its use in regression.
2. **Overfitting & Regularization:** Discussion of overfitting in complex models and the introduction of regularization (e.g., L2 regularization).
3. **Basic Probability and Cost Functions:** Overview of cost functions (such as Mean Squared Error) used in regression.

This Lecture Focuses on:

1. The formulation and intuition behind Logistic Regression.
2. Transforming linear outputs to probabilities using the Sigmoid Function.
3. Defining the Decision Boundary.
4. Deriving the Cross-Entropy (Log Loss) Cost Function from maximum likelihood (ESL 4.4.1, PML 10.2.1–10.2.2).
5. Gradient Descent.
6. Incorporating Regularization.
7. Extending to Multi-class Classification using the Softmax function and the $argmax$ decision rule.

1 Introduction

Logistic Regression is a supervised learning algorithm used for classification. Unlike linear regression—which outputs continuous values—logistic regression estimates the probability that an input x belongs to a particular class (usually labeled 0 or 1). The probability is then thresholded to obtain a discrete classification.
Example: In an email spam detection system, features (such as word frequency and email length) are used to compute the probability that an email is spam (class 1) or not spam (class 0).

2 Linear Model and Hypothesis

The model begins by computing a linear combination of the input features:

$$z = w^T x + b,$$

where:

- $x \in \mathbb{R}^n$ is the feature vector,
- $w \in \mathbb{R}^n$ is the weight vector,
- $b \in \mathbb{R}$ is the bias term.

This linear model is the foundation for the logistic hypothesis.

3 The Sigmoid Function

To transform the linear output z into a probability, we apply the sigmoid function.

3.1 Definition

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

Thus, the logistic regression hypothesis is:

$$h_\theta(x) = \sigma(w^T x + b) = \frac{1}{1 + e^{-(w^T x + b)}}.$$

3.2 Properties and Derivative

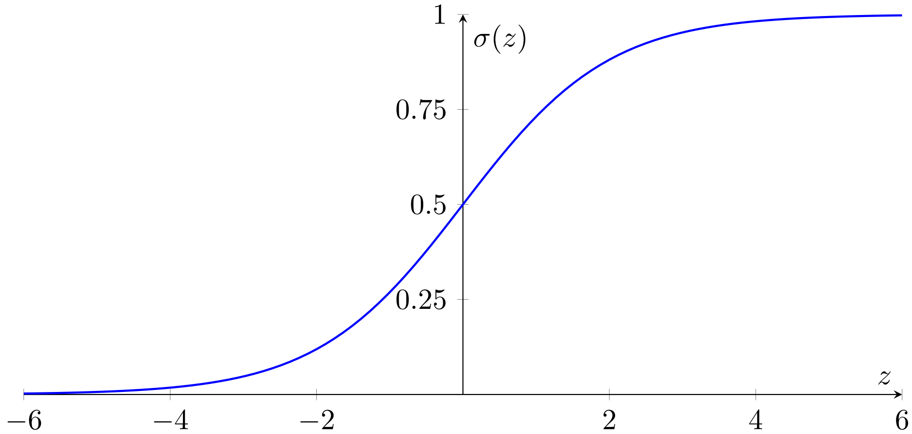
- As $z \to +\infty$, $\sigma(z) \to 1$; as $z \to -\infty$, $\sigma(z) \to 0$.
- $\sigma(0) = 0.5$ which defines the natural threshold.
- The derivative is:

$$\sigma'(z) = \sigma(z)\big(1 - \sigma(z)\big).$$

- **Example:** For $z = 2$:

$$\sigma(2) \approx \frac{1}{1 + e^{-2}} \approx 0.88.$$

3.3 standard sigmoid plot



4 Decision Boundary

After computing $h_\theta(x)$, the output probability is converted to a class label by thresholding.

4.1 Classification Rule

$$\hat{y} = \begin{cases} 1, & \text{if } h_\theta(x) \geq 0.5, \\ 0, & \text{if } h_\theta(x) < 0.5. \end{cases}$$

Since $\sigma(0) = 0.5$, the decision boundary is:

$$w^T x + b = 0.$$

4.2 Example

If $w^T x + b = 0.8$, then:

$$\sigma(0.8) \approx 0.69 \quad \text{(classified as 1)}.$$

If $w^T x + b = -1.2$, then:

$$\sigma(-1.2) \approx 0.23 \quad \text{(classified as 0)}.$$

5 Cost Function: Cross-Entropy Loss

The model is trained by minimizing the cross-entropy loss, derived from the maximum likelihood principle (see ESL 4.4.1 and PML 10.2.1–10.2.2).

5.1 Loss for a Single Example

For an example $(x^{(i)}, y^{(i)})$ with $y^{(i)} \in \{0, 1\}$ and $\hat{y}^{(i)} = h_\theta(x^{(i)})$:

$$\text{cost}\big(h_\theta(x^{(i)}), y^{(i)}\big) = -\Big[y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log\big(1 - \hat{y}^{(i)}\big)\Big].$$

5.2 Overall Cost Function

Averaging over m examples:

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

5.3 Examples

- If $y = 1$ and $h_\theta(x) = 0.9$, then cost $\approx -\log(0.9) \approx 0.105$.
- If $y = 1$ but $h_\theta(x) = 0.2$, cost $\approx -\log(0.2) \approx 1.609$.
- If $y = 0$ and $h_\theta(x) = 0.1$, cost $\approx -\log(0.9) \approx 0.105$.
- If $y = 0$ but $h_\theta(x) = 0.8$, cost $\approx -\log(0.2) \approx 1.609$.

6 Gradient Descent

The cost function $J(\theta)$ is minimized using gradient descent.

6.1 Gradient Computation

For each weight w_j:

$$\frac{\partial J(\theta)}{\partial w_j} = \frac{1}{m}\sum_{i=1}^{m}\left[\sigma(w^T x^{(i)} + b) - y^{(i)}\right]x_j^{(i)}.$$

For the bias b:

$$\frac{\partial J(\theta)}{\partial b} = \frac{1}{m}\sum_{i=1}^{m}\left[\sigma(w^T x^{(i)} + b) - y^{(i)}\right].$$

7 Regularization

To prevent overfitting, a regularization term is added.

7.1 L2 Regularization

The regularized cost function is:

$$J_{\text{reg}}(\theta) = J(\theta) + \frac{\lambda}{2m}\sum_{j=1}^{n} w_j^2,$$

where λ is the regularization parameter.
Example: With $w = [0.5645, -0.2785]$ and $\lambda = 0.1$ (assuming $m = 1$), the penalty is approximately:

$$\frac{0.1}{2}\left(0.5645^2 + (-0.2785)^2\right) \approx 0.0396.$$

8 Multi-class Logistic Regression (Softmax Regression)

For problems with more than two classes, logistic regression is extended via the softmax function.

8.1 Softmax Function and Hypothesis

For K classes, compute for each class k:

$$z_k = w_k^T x + b_k, \quad k = 1, \ldots, K.$$

Then the probability that x belongs to class k is:

$$P(y = k \mid x; \theta) = \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}}.$$

8.2 Decision Rule

The predicted class is determined by:

$$\hat{y} = \arg\max_k \, P(y = k \mid x; \theta).$$

8.3 Multi-class Cost Function

The cross-entropy loss for multi-class classification is:

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \sum_{k=1}^{K} 1\{y^{(i)} = k\} \log P(y = k \mid x^{(i)}; \theta).$$

8.4 Example: Digit Classification

Assume an image classifier outputs scores for three classes (digits 0, 1, 2):

$$z = [2.0,\ 1.0,\ 0.1].$$

Then:

$$\begin{aligned} e^{2.0} &\approx 7.39, \\ e^{1.0} &\approx 2.72, \\ e^{0.1} &\approx 1.105. \end{aligned}$$

The sum is approximately 11.215, so:

$$P(y = 0 \mid x) \approx \frac{7.39}{11.215} \approx 0.66, \quad P(y = 1 \mid x) \approx \frac{2.72}{11.215} \approx 0.242, \quad P(y = 2 \mid x) \approx \frac{1.105}{11.215} \approx 0.0985.$$

Thus, the predicted label is:

$$\hat{y} = \arg\max_k \, P(y = k \mid x; \theta) = 0.$$

9 Summary and Conclusions

In Lecture 4, we covered the following:

- The linear model $z = w^T x + b$ serves as the basis of the hypothesis.
- The sigmoid function, $\sigma(z) = \frac{1}{1+e^{-z}}$, transforms z into a probability.
- The decision boundary is defined by $w^T x + b = 0$, with classification performed by thresholding at 0.5.
- The cross-entropy loss function,

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} \log\left(h_\theta(x^{(i)})\right) + (1 - y^{(i)}) \log\left(1 - h_\theta(x^{(i)})\right) \right],$$

  is derived via maximum likelihood.
- L2 regularization is incorporated as:

$$\frac{\lambda}{2m} \sum_{j=1}^{n} w_j^2,$$

  to control overfitting.
- For multi-class classification, the softmax function extends the model:

$$P(y = k \mid x; \theta) = \frac{e^{w_k^T x + b_k}}{\sum_{j=1}^{K} e^{w_j^T x + b_j}},$$

  with prediction via

$$\hat{y} = \arg\max_k \; P(y = k \mid x; \theta).$$

10 References

- *The Elements of Statistical Learning* (ESL), Section 4.4.1.
- *Probabilistic Machine Learning: An Introduction* (PML), Sections 10.2.1 and 10.2.2.
- *Lec-4 class notes*