COMP24112: Machine Learning

Chapter 5: Loss Functions III

Dr. Tingting Mu

Email: tingting.mu@manchester.ac.uk

Content

- Typical approaches of constructing losses for classification
  - Non-probabilistic (Part B)
  - **Probabilistic (Part C)**

- ***Cross entropy loss based on class posterior.***

$$p(\text{class } k \mid \mathbf{x})$$

Cross Entropy

- **Cross entropy** measures distance between probability distributions.

- Its discrete version can be used to examine the **distance** between the predicted class probabilities (**posterior**) and the **true probabilities**.

$$H(p,q) = -\left[p(1)\log\left(q(1)\right) + p(0)\log\left(q(0)\right)\right] \text{ (two classes)}$$

$$H(p,q) = -\left[p(1)\log\left(q(1)\right) + p(2)\log\left(q(2)\right) + \cdots + p(c)\log\left(q(c)\right)\right] \text{ (multiple classes)}$$

Cross Entropy Loss

- **Binary classification**: $H(p,q) = -\left[p(1)\log\left(q(1)\right) + p(0)\log\left(q(0)\right)\right]$

- **0/1 label coding** for a sample (***x***, *y*).
  - If y=1, x is from class 1, which means *p(1) =100%* and *p(0) =0%.*
  - If y=0, x is from class 0, which means *p(1) =0%* and *p(0) = 100%.*
  - Therefore, ***p(1) =y*** and ***p(0) =1-y***.

- Cross entropy loss computed over N training samples is

$$O = -\sum_{i=1}^{N}\left[y_i \log_b\left(p\left(c_1 \middle| \mathbf{x}_i\right)\right) + \left(1 - y_i\right)\log_b\left(p\left(c_2 \middle| \mathbf{x}_i\right)\right)\right]$$

You can use natural log (*ln*, *b*=*e*) or log base 2 (*b*=2).

Cross Entropy Loss

- **Multi-class classification**

y_{ik}*=1* means that the i-th sample belongs to class k, y_{ik}*=0* otherwise.

  - 1-of-K label coding scheme: $\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1c} \\ y_{21} & y_{22} & \cdots & y_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ y_{N1} & y_{N2} & \cdots & y_{Nc} \end{bmatrix}$, where $y_{ik} \in \{0,1\}$

  - Cross entropy loss computed over N training samples is

$$O = -\sum_{i=1}^{N}\sum_{k=1}^{c} y_{ik} \log_b\left(p\left(c_k \middle| \mathbf{x}_i\right)\right)$$

Cross Entropy Loss

- Different models give you different ways to formulate $p\left(c_k \middle| \mathbf{x}\right)$.

- Logistic regression: A linear classification model trained using cross-entropy loss.

- **Classification Losses based on likelihood.**

Likelihood Maximization for Classification

- One way to train the model is to **maximise the likelihood (or log likelihood) function**.
  - Maximum likelihood estimator (MLE):

$$\max_{\theta} p(\text{data} \mid \theta)$$

  - Often it is more convenient to use log likelihood:

$$\max_{\theta} \log p(\text{data} \mid \theta)$$

Assumption on Class Label Distribution

- **Binary classification**: Assume the class label follows Bernoulli distribution.

$$p(y|\theta) = \theta^{y}(1-\theta)^{1-y} = \begin{cases} \theta, & \text{if } y = 1, \\ 1-\theta, & \text{if } y = 0. \end{cases}$$

- **Multi-class classification**: Assume the class label follows categorical (multinomial) distribution.

$$p(\mathbf{y}|\theta_1, \theta_2, \ldots \theta_c) = \prod_{k=1}^{c} \theta_k^{y_k}$$

1-of-K coding scheme:

$$\mathbf{y} = [y_1, y_2, \ldots y_k]^T$$

$y_k = 1$ if the sample is from class k

$y_k = 0$ otherwise

Likelihood of An Individual Sample

Consider the *i*-th training sample $(\mathbf{x}_i, y_i)$

- **Binary classification**:

$$p(\mathbf{x}_i, y_i \,|\, \theta) = \theta(\mathbf{x}_i)^{y_i}\big(1 - \theta(\mathbf{x}_i)\big)^{1-y_i}$$

- **Multi-class classification**: Assume the class label follows categorical (multinomial) distribution (generalise Bernoulli distribution to more than two options):

$$p\left(\mathbf{x}_i, \mathbf{y}_i \,|\, \theta\right) = \prod_{k=1}^{c} \theta_k(\mathbf{x}_i)^{y_{ik}}$$

Likelihood of N Training Samples

Given N training samples and assume sample independence.

- **Binary classification**:

$$L = \prod_{i=1}^{N} p(\mathbf{x}_i, y_i \mid \theta) = \prod_{i=1}^{N} \theta(\mathbf{x}_i)^{y_i}\left(1 - \theta(\mathbf{x}_i)\right)^{1-y_i}$$

- **Multi-class classification**:

$$L = \prod_{i=1}^{N} p\left(\mathbf{x}_i, \mathbf{y}_i \mid \theta\right) = \prod_{i=1}^{N}\prod_{k=1}^{c} \theta_k(\mathbf{x}_i)^{y_{ik}}$$

Example

- You can **model your θ using a linear model**.
  - Binary classification: Apply logistic sigmoid function to a linear model.

$$\theta(\mathbf{x}) = \frac{1}{1+\exp\left(-\mathbf{w}^T\tilde{\mathbf{x}}\right)}$$

  - Multi-class classification: Apply softmax function to a linear model.

$$\theta_k(\mathbf{x}) = \frac{\exp\left(\mathbf{w}_k^T\tilde{\mathbf{x}}\right)}{\sum_{j=1}^{c}\exp\left(\mathbf{w}_j^T\tilde{\mathbf{x}}\right)} \quad ,k = 1,2,\ldots c$$

This gives you again logistic regression.

Negative Log likelihood Loss

- Classification loss: negative log likelihood.
- Negative log likelihood loss is equal to the cross-entropy loss, if

$$\theta(\mathbf{x}) = p(c_1|\mathbf{x})$$

$$\theta_k(\mathbf{x}) = p(c_k|\mathbf{x})$$

Example1: Cross-Entropy Loss for Binary Classification

- A logistic regression model returns the following posterior class probabilities for the 3 samples as below:

A	y	$p(c_1\|x)$	$p(c_2\|x)$
x_1	1	0.8	0.2
x_2	1	0.9	0.1
x_3	0	0.2	0.8

$$O = -\sum_{i=1}^{N}\left[y_i \log_b\left(p\left(c_1 \middle| \mathbf{x}_i\right)\right) + \left(1-y_i\right)\log_b\left(p\left(c_2 \middle| \mathbf{x}_i\right)\right)\right]$$

- Compute this model's cross-entropy loss using these 3 samples.

$$O_A = -\left(1\times\ln 0.8 + 0\times\ln 0.2\right) - \left(1\times\ln 0.9 + 0\times\log 0.1\right) - \left(0\times\ln 0.2 + 1\times\ln 0.8\right)$$
$$= -\left(\ln 0.8 + \ln 0.9 + \ln 0.8\right) = 0.55$$

The University of Manchester

Example2:
Negative log likelihood Loss for Multi-class Classification

- A 3-class classification model is trained by MLE assuming categorical distribution. The ground truth labels and estimated theta function for the following 4 samples are provided:

	y	$\Theta_1(x)$	$\Theta_2(x)$	$\Theta_3(x)$
x_1	1	0.7	0.2	0.1
x_2	3	0.5	0.3	0.2
x_3	3	0.1	0.1	0.8
x_4	2	0.3	0.6	0.1

$$L = \prod_{i=1}^{N} \prod_{k=1}^{c} \theta_k(\mathbf{x}_i)^{y_{ik}}$$

- Compute this model's Negative log likelihood loss using these 4 samples.

$$L = \left(0.7^1 \times 0.2^0 \times 0.1^0\right) \times \left(0.5^0 \times 0.3^0 \times 0.2^1\right) \times \left(0.1^0 \times 0.1^0 \times 0.8^1\right) \times \left(0.3^0 \times 0.6^1 \times 0.1^0\right)$$

$$= 0.7 \times 0.2 \times 0.8 \times 0.6 = 0.0672$$

$$-\ln(L) = 2.7$$

Chapter 5 Summary: A, B and C

- Regression losses:
  - Sum of squares error
  - Mean squared error
- Classification losses:
  - Sum of squares error
  - Hinge loss
  - Cross entropy loss
  - Likelihood and log likelihood based
- Linear least squares (LLS) approach for classification and regression
- Regularization, regularized LLS