cs577 Assignment 2: Solution

Yuanxing Cheng, A20453410, CS577-f22
Department of Mathematics
Illinois Institute of Technology

October 26, 2022

Theoretical questions

Loss

1

> Write the equation for L1, L2, Huber and Log-cosh loss function and compare them. Explain their advantages/purposes.

Let $d_j = \hat{y}_j^{(i)} - y_j^{(i)}$ for i-th sample at j-th dimension, we have

- L1 loss: $\sum_j |d_j|$
- L2 loss: $\sum_j d_j^2$
- Huber loss: $\sum_j \rho_\sigma(d_j)$ where $\rho_\sigma(d) = \begin{cases} \frac{1}{2}d^2, & |d| \le \sigma \\ \sigma(d - \frac{1}{2}\sigma), & \text{otherwise} \end{cases}$
- Log-cosh loss: $\sum_j \log(\cosh(d_j))$

Their advantages/purposes are:

- L1 loss: use absolute value to avoid cancelling of positive and negative d_j
- L2 loss: use square to avoid this cancelling while make loss differentiable
- Huber loss: to cap the loss for outliers
- log-cosh: cap the loss for outliers and differentiable

2

> Write the equation for cross-entropy loss and explain how it is derived using maximum log-likelihood. Explain the worst cross-entropy value you expect for random assignment.

$$\hat{y}_j^{(i)} = P(y = j \mid X^{(i)})$$

then the negative log likelihood is

$$l(\theta) = -\log\left(\prod_{i=1}^{m}\prod_{j=1}^{k} P(y = j \mid X^{(i)})^{y_j^{(i)}}\right) = -\sum_{i=1}^{m}\sum_{j=1}^{k} y_j^{(i)} \log\left(\hat{y}_j^{(i)}\right)$$

The random assignment will lead to $-\log k$.

3

Write the equation for softmax loss and describe when to use it

$$l(\theta) = -\sum_{i=1}^{m}\sum_{j=1}^{k} y_j^{(i)} \log\left(\hat{y}_j^{(i)}\right) = -\sum_{i=1}^{m}\sum_{j=1}^{k} y_j^{(i)} \left(z_j - \log\sum_{j=1}^{k}\exp(z_j)\right)$$

When to use: multi-class classification problem and when using softmax as activation before output layer

4

Write the equation for Kullback-Liebler loss and explain its meaning. Explain the circumstances under which there is no difference between using cross-entropy or Kullback-Liebler to train the network.

$$l(\theta) = \sum_{i=1}^{m}\sum_{j=1}^{k} y_j^{(i)} \log\left(\frac{y_j^i}{\hat{y}_j^{(i)}}\right)$$

meaning: the similarity between two distributions. This equals to cross-entropy loss when $\sum_{i=1}^{m}\sum_{j=1}^{k} y_j^{(i)} \log\left(y_j^i\right) = 0$

5

Explain the Hinge loss and squared Hinge loss. Describe the fundamental idea behind it and the worst value you expect for it before learning.

Hinge loss: used in binary classification while the labels are $-1, 1$. It's positive if prediction and input disagree in sign or if agree in sign and $\left|\hat{y}^{(i)}\right| < \sigma$. It's negative if predictions agree in sign and $\left|\hat{y}^{(i)}\right| > \sigma$. It has the form:

$$l(\theta) = \sum_{j} \max\left(0, \sigma - y_j^{(i)}\hat{y}_j^{(i)}\right)$$

The squared hinge loss also measure the label distance with margin σ. It has the form:

$$l(\theta) = \frac{1}{2}\sum_{j=1}^{k} \max\left(0, \hat{y}_j^{(i)} - \hat{y}_{true}^{(i)} + \sigma\right)^2$$

The worst value can go infinity large.

6

Compute the hinge loss for a 3-class classification problem with three examples with label $1, 2, 3$, with prediction scores $\hat{y}^{(1)} = (0.5, 0.4, 0.3)$, $\hat{y}^{(2)} = (1.3, 0.8, -0.6)$ and $\hat{y}^{(1)} = (1.4, -0.4, 2.7)$ (one-against-all-others).

$$\begin{aligned}
L_1 &= \max(0, 0.4 - 0.5 + 1) + \max(0, 0.3 - 0.5 + 1) = 1.7\\
L_2 &= \max(0, 1.3 - 0.8 + 1) + \max(0, -0.6 - 0.8 + 1) = 1.5\\
L_3 &= \max(0, 1.4 - 2.7 + 1) + \max(0, -0.4 - 2.7 + 1) = 0
\end{aligned}$$

7

> Explain the purpose of adding a regularization term to the loss function. Explain the difference between L1 and L2 regularization and how they affect the weight distribution in the network. Explain the way to choose the regularization term coefficient.
>
> ---
>
> Purpose:
>
> - simple explanation is better (for better generalization)
> - when parameters are small enough, we can remove it and make the model simpler
> - smaller parameters are more stable that will generalize better
>
> Difference between L1 and L2: L2 is more sensitive to outliers due to the square. Method to choose regularization coefficient is to plot the path of the coefficient λ vs the model parameters.

8

> Explain how L1 and L2 loss terms affect gradients in the network
>
> ---
>
> As regularization term is added into the loss, the gradients also changes. It slightly alters the coefficients by a linear function on λ and thus prevent overfitting.

9

> Explain the difference between kernel, bias, and activity regularization
>
> ---
>
> - kernel regularizer: on model parameters θ
> - bias regularizer: on θ_0
> - activity regularizer: on $\hat{y}$

Regularization

1

> Explain how weight decay is related to adding a regularization term to the loss function.
>
> ---
>
> Add the regularization term results in extra term in the gradient and thus in gradient descent method, extra term is substructed and that's weight decay.

2

> Explain how early stopping to prevent overfitting is performed. Explain the strategies to reuse the validation data.
>
> ---
>
> Early stopp when validation error starts to increase or when training error stops decreasing and this prevent overfitting. To reuse validation data, we have 2 strategies.
>
> - use only train data and then train for number of iterations determine from validation data
> - after early stop, continue trainnnning from previous weights with train data while validation loss is bigger than training loss

3

Explain how data augmentation is performed and how it assists in prevent overfitting.

Add synthetic data to increase variability in training so that we have better generalization

- augment in feature or data domain
- augment by interpolating between examples on by adding noise
- augment by transforming img data: crop, rotate, rescale, intensity
- other popular method in image classification, that are illumination/rotation/scale invariante

4

Explain how dropout is performed. What are advantages/disadvantages of dropout?

Steps for drop out.

1. at each training stage, drop out units in fully connected layers with probability $1 - \rho$ where ρ is a hyper-parameter.
2. removed nodes are reinstated with original weights in the subsequent state

Advantages:

- reduce node interations
- reduce overfitting
- increase training speed
- reduce dependency on a single node
- disttribute features across multiple nodes

Disadvantages: longer training time

5

Explain how the expected value of all combinations of dropped out networks can be approximated efficiently during testing.

At testing, we can multipy the output of each node by dropout probability so that it equals to the expected value.

6

Explain how batch normalization is performed during training and during testing. In what way does batch normalization introduces randomness into training.

During training the layer output are normalized using output mean μ_j and output standard deviation σ_j: $\hat{z}_j^{(i)} = \frac{z_j^{(i)} - \mu_j}{\sigma_j}$. These numbers are stored and used again in testing process. During training, because batches are randomly selected, batch normalization adds randomness into the training and thus reduces overfitting.

7

Explain the purpose of scale and shift parameters in batch normalization. What are the values of scale and shift parameters that will cause the normalization to be canceled? Explain how the scale and shift parameters can be learned and what is a good initial value for them.

To terminate training, we need scale and shift after normalization: $\tilde{z}_j = r_j \hat{z}_j + \beta_j$. If $r_j = \sigma_j$ and $\beta_j = \mu_j$, batch normalization is canceled. These values can be learned if batch normalization is not needed. A good initial values could be zero for mean and one for standard deviation.

8

Explain how ensemble classifiers can assist with overfitting. Describe the possible strategies for producing ensemble classifiers.

Ensemble classifiers train multiple independent models with:

- change data
- change parameters
- record multiple snapshots of the model during training with various learning rate