Introduction to Machine Learning

General Note About Linear Classifiers

Mingchen Gao

Outline

Contents

1 Linear Classifers and Loss Function **1**
1.1 Regularizers 3
1.2 Approximate Regularization 4

1 Linear Classifers and Loss Function

- Linear binary classification can be written as a general optimization problem:

$$\min_{\mathbf{w},b} L(\mathbf{w}, b) = \min_{\mathbf{w},b} \sum_{n=1}^{N} \mathbb{I}(y_n(\mathbf{w}^\top \mathbf{x}_n + b) < 0) + \lambda R(\mathbf{w}, b)$$

- $\mathbb{I}$ is an **indicator function** (1 if (.) is negative, 0 otherwise)
- Objective function = **Loss function** + λ**Regularizer**
- Objective function wants to **fit training data well** and **have simpler solution**
- Combinatorial optimization problem
- NP-hard
- No polynomial time algorithm

- Loss function is non-smooth, non-convex
- Small changes in $\mathbf{w}, b$ can change the loss by lot
- Different linear classifiers use different approximations to 0-1 loss
  - Also known as *surrogate loss functions*

Support Vector Machines

- **Hinge Loss**

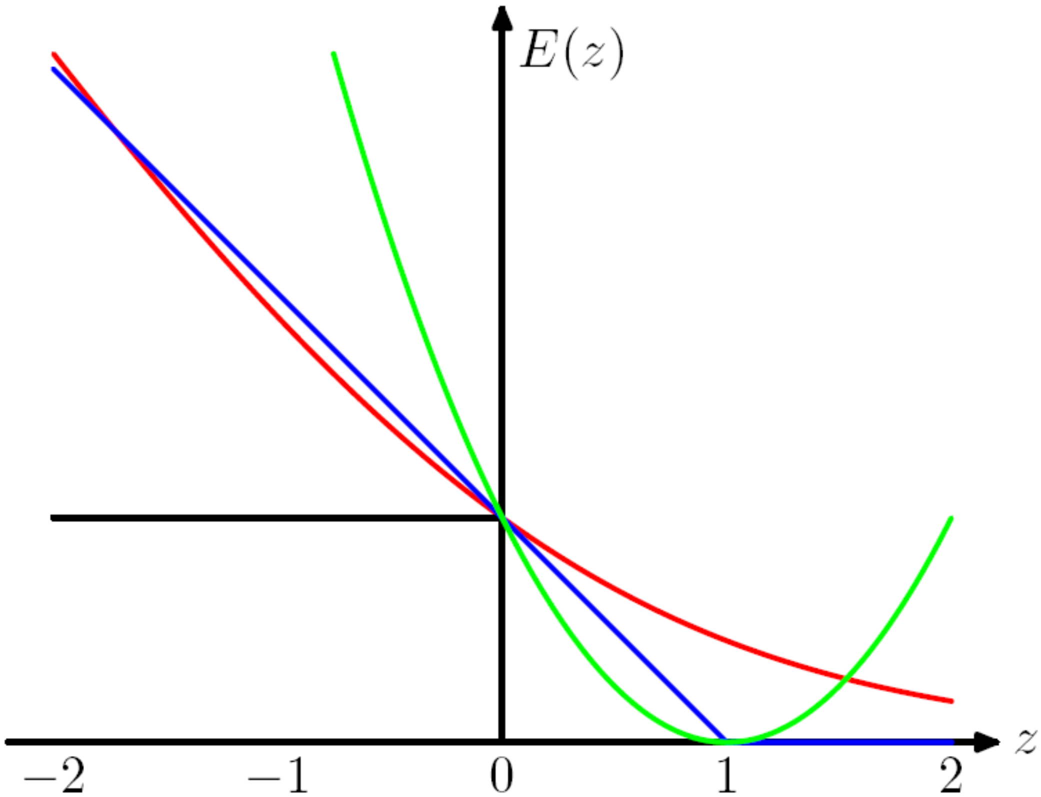
Squared Loss

- **Squared Loss**

Logistic Regression

- **Log Loss**

- black, indicator loss
- green, squared loss
- red, log loss
- blue, hinge loss



1.1 Regularizers

- Recall the optimization problem for linear classification

$$\min_{\mathbf{w},b} L(\mathbf{w}, b) = \min_{\mathbf{w},b} \sum_{n=1}^{N} \mathbb{I}(y_n(\mathbf{w}^\top \mathbf{x}_n + b) < 0) + \lambda R(\mathbf{w}, b)$$

- What is the role of the regularizer term?
  - Ensure simplicity
- Ideally we want most entries of $\mathbf{w}$ to be zero
- Why?

- Desired minimization

$$R(\mathbf{w}, b) = \sum_{d=1}^{D} \mathbb{I}(w_d \neq 0)$$

- NP Hard

The reason we want most entries in the weight vector $\mathbf{w}$ to be 0 is so that the prediction depends only on a few features. This would ensure that changes in x_d for those features will not change the prediction, hence higher bias.

1.2 Approximate Regularization

- **Norm based regularization**
  - l_2 squared norm
    $$\|\mathbf{w}\|_2^2 = \sum_{d=1}^{D} w_d^2$$
  - l_1 norm
    $$\|\mathbf{w}\|_1 = \sum_{d=1}^{D} |w_d|$$
  - l_p norm
    $$\|\mathbf{w}\|_p = (\sum_{d=1}^{D} w_d^p)^{1/p}$$
  - Norm becomes non-convex for $p < 1$
  - l_1 norm gives best results
  - l_2 norm is easiest to deal with