Lecture Notes for Deep Learning

1 Introduction

m	number of samples
n	number of features
y	actual label
$\hat{y}$	predicted label
ℓ	loss function
ς	activation function

The training set is a set of samples used for “building” or “finding” a model. The test set is a set of samples used for “evaluating” or “checking” a model. A metric is either a loss function or testing performance measure.[1]

Informally, a model is said to be overfit if its capacity is too high, and underfit if it is too low. Regularization mitigates overfitting by reducing the Vapnik–Chervonenkis (VC) dimension of the hypothesis class of the model. The ridge and lasso regularization terms “encourage” smaller weights quadratically and linearly, respectively; the latter allows weights to become zero.[2]

For brevity, we use bold-faced function symbols to denote an element-wise application of a function to a vector or matrix.

[1] A “feature” in machine learning is said to correspond to an “independent variable” in statistics.

[2] In describing (the hypothesis class of) a model, “capacity” is synonymous with “complexity”, “expressiveness”, and “flexibility”.

2 Linear Regression and Logistic Binary Classification

Definition 1 (MSE metric)**.** Let $\boldsymbol{y}, \hat{\boldsymbol{y}} \in \mathbb{R}^m$.

$$\mathsf{MSE}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{m}\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2$$

Definition 2 (Linear regressor)**.** Let $\boldsymbol{X} \in \mathbb{R}^{m\times(n+1)}$, and $\boldsymbol{\beta} \in \mathbb{R}^{n+1}$.

$$\hat{\boldsymbol{y}}(\boldsymbol{X}; \boldsymbol{\beta}) = \boldsymbol{X\beta}$$

Theorem 1. Let $\boldsymbol{X} \in \mathbb{R}^{m\times(n+1)}$, and $\boldsymbol{\beta} \in \mathbb{R}^{n+1}$.

$$\operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^{n+1}} \mathsf{MSE}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}\boldsymbol{Xy}$$

Definition 3 (Sigmoid function)**.**

$$\sigma : \mathbb{R} \to \mathbb{R}_{(0,1)}$$

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

Definition 4 (Logistic binary classifier)**.** Let $\boldsymbol{X} \in \mathbb{R}^{m\times n}$, and $\boldsymbol{w} \in \mathbb{R}^n$.

$$\hat{\boldsymbol{y}}(\boldsymbol{x}; \boldsymbol{w}) = \boldsymbol{\sigma}(\boldsymbol{Xw})$$

Remark 1. Each entry can be seen as how probable its corresponding input entry will be in the category associated with “1”.

Definition 5 (Cross-entropy loss function)**.** Let $\boldsymbol{y}, \hat{\boldsymbol{y}} \in \mathbb{R}^m$.

$$\begin{aligned}\mathsf{CE}(\boldsymbol{y}, \hat{\boldsymbol{y}}) &= -\frac{1}{m}\sum_{i=1}^{m} y_i \ln(\hat{y}_i) + (1 - y_i)\ln(1 - \hat{y}_i) \\ &= -\frac{1}{m}(\boldsymbol{y} \cdot \mathbf{ln}(\hat{\boldsymbol{y}}) + (\mathbf{1} - \boldsymbol{y}) \cdot \mathbf{ln}(\mathbf{1} - \hat{\boldsymbol{y}}))\end{aligned}$$

3 Multiperceptron Regression

4 Appendix

Proposition 1. Let $x \in \mathbb{R}$.

$$0 < \sigma(x) < 0.5 \Leftrightarrow x < 0$$

$$\sigma(x) = 0.5 \Leftrightarrow x = 0$$

$$0.5 < \sigma(x) < 1 \Leftrightarrow x > 0$$