TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

7.2 Loss View

Machine Learning 1: Foundations

Marius Kloft *(TUK)*

Recall:

SVM:

$$\min_{b,\mathbf{w}} \ \frac{1}{2}\|\mathbf{w}\|^2 \ + \ C\sum_{i=1}^{n} \max\Big(0, 1 - y_i(\langle \mathbf{w}, \underset{\substack{\uparrow \\ \text{kernel feature map of a kernel } k}}{\phi_k}(\mathbf{x}_i)\rangle + b)\Big)$$

Logistic Regression (LR):

$$\min_{b,\mathbf{w}} \ \frac{1}{2}\|\mathbf{w}\|^2 \ + \ C\sum_{i=1}^{n} \ln\Big(1 + \exp(-y_i(\langle \mathbf{w}, \phi_k(\mathbf{x}_i)\rangle + b))\Big)$$

ANN:

$$\min_{b,\mathbf{w},W} \ \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{2}\sum_{l=1}^{L}\|W_l\|_{\text{Fro}}^2 + C\sum_{i=1}^{n} \ln(1+\exp(-y_i(\mathbf{w}^\top \phi_W(\mathbf{x}_i)+b)))$$

All these methods can be unified into a single equation.

For simplicity, we had introduced ANNs without bias b in the ANN class. Here we use a bias b, which makes a lot sense, for the same reasons as it makes sense also in SVMs and LR.

Unifying View

Unifying formulation of linear, kernel, and neural learning

$$\min_{[W,]\ b,\mathbf{w}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\ell(y_i(\langle\mathbf{w},\phi(\mathbf{x}_i)\rangle + b)) \quad \left[+\frac{1}{2}\sum_{l=1}^{L}\|W_l\|_{\text{Fro}}^2\right],$$

where

- $\ell(t) := \max(0, 1-t)$ for SVM ("hinge loss")
- $\ell(t) := \ln(1+\exp(-t))$ for LR and ANN ("logistic loss")

and

- $\phi :=$ id for linear SVM and linear LR
- $\phi := \phi_k$ for kernel SVM and kernel LR
- $\phi := \phi_W$ for ANN.

The terms in brackets apply only for ANNs.

For a test point $\mathbf{x}$, we predict $f(\mathbf{x}) := \text{sign}(\langle\mathbf{w},\phi(\mathbf{x})\rangle + b)$.

Five Popular Learning Machines in One Equation

The following table summarizes the result of the previous slide:

loss ℓ \ map ϕ	id	ϕ_k	ϕ_W
hinge	linear SVM	kernel SVM	—[1]
logistic	linear LR	kernel LR	ANN

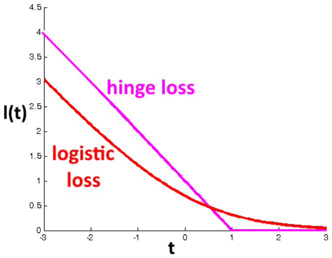
[1] The hinge loss is theoretically possible but uncommon in neural networks.

The Loss

The unifying equation contains,
for every training example $(\mathbf{x}_i, y_i)$,
a term

$$\ell(\underbrace{y_i \langle \mathbf{w}, \phi(\mathbf{x}) \rangle + b}_{=:t_i}),$$

the "loss" of the ith example.



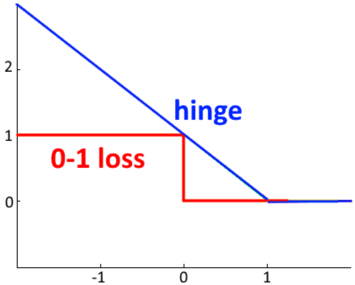
Interpretation:

- In the unifying formulation, we minimize the loss $\sum_i \ell(t_i)$
- This promotes solutions where t_i is large
- Make sense: we want t_i being large
  - because $t_i > 0$ means the ith label is correctly predicted

Understanding the Loss

Definition

The function $\ell(t) := \mathbf{1}_{t<0}$ is called **0-1 loss**. In this context, $\mathbf{1}$ denotes the indicator function.



Observations:

- The 0-1 loss is one when the label is misclassified ($t < 0$) and zero otherwise
- Thus the cumulative 0-1 loss $\sum_{i=1}^{n} \ell\left(y_i\left(\langle \mathbf{w}, \phi(\mathbf{x}_i)\rangle + b\right)\right)$ measures the number of training errors
- Makes sense to get this error as small as possible
- Turns out to be an NP hard problem! :(

Idea in SVM, LR, and ANN: replace the difficult 0-1 loss by a convex approximation—the hinge or logistic loss