8.4 Unifying View

Machine Learning 1: Foundations

Marius Kloft *(TUK)*

Unifying View

Recall our unifying view

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

which comprises (linear and kernelized) SVM and LR, as well as ANN in just one equation.

Unifying View

Recall our unifying view

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

which comprises (linear and kernelized) SVM and LR,
as well as ANN in just one equation.

Wouldn't it be nice to add also regression into this equation? :)

Unifying View

Recall our unifying view

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

which comprises (linear and kernelized) SVM and LR, as well as ANN in just one equation.

Wouldn't it be nice to add also regression into this equation? :)

In order to do so, we slightly change our notation of the loss:

$$l(t, y) := \begin{cases} (t - y)^2 & \text{for regression} \\ \ell(yt) & \text{for classification} \end{cases}$$

Unifying View

Recall our unifying view

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

which comprises (linear and kernelized) SVM and LR, as well as ANN in just one equation.

Wouldn't it be nice to add also regression into this equation? :)

In order to do so, we slightly change our notation of the loss:

$$l(t,y) := \begin{cases} (t-y)^2 & \text{for} \quad \text{regression} \\ \ell(yt) & \text{for} \quad \text{classification} \end{cases}$$

We obtain...

Unifying View of Regression and Classification

Unifying formulation of linear, kernel, and neural classification and regression

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

where

- $l(t,y) := \max(0, 1-yt)$ for SVM ("hinge loss")
- $l(t,y) := \ln(1+\exp(-yt))$ for LR and ANN ("logistic loss")
- $l(t,y) := (t-y)^2$ for regression

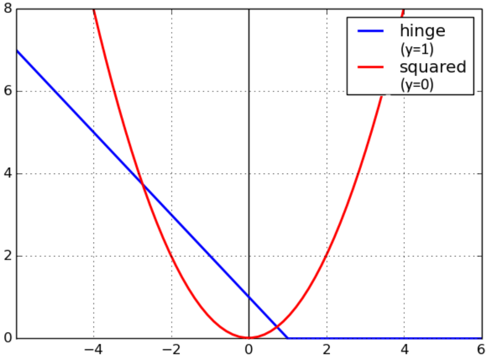
and

- $\phi := \text{id}$ for linear SVM, linear LR, and RR
- $\phi := \phi_k$ for kernel SVM, kernel LR, and KRR
- $\phi := \phi_W$ for ANN and DR.

The terms in brackets apply only to ANN and DR.

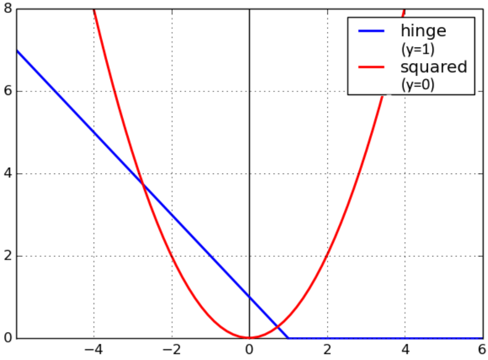
Unifying View Reveals: Classification and Regression Differ only in the Loss

- Regression uses the squared loss: $l(t, y) = (t - y)^2$
- E.g., the SVM uses the hinge loss: $l(t, y) = \max(0, 1 - yt)$



Unifying View Reveals: Classification and Regression Differ only in the Loss

- Regression uses the squared loss: $l(t, y) = (t - y)^2$
- E.g., the SVM uses the hinge loss: $l(t, y) = \max(0, 1 - yt)$



Can we derive a SVM-style regression method?

Support Vector Regression (SVR)

SVR uses the same regularizer as SVM, but the following loss:

Definition

The ϵ**-insensitive loss** is defined as
$\ell(t, y) := \max(0, |y - t| - \epsilon)$

Support Vector Regression (SVR)

SVR uses the same regularizer as SVM, but the following loss:

Definition

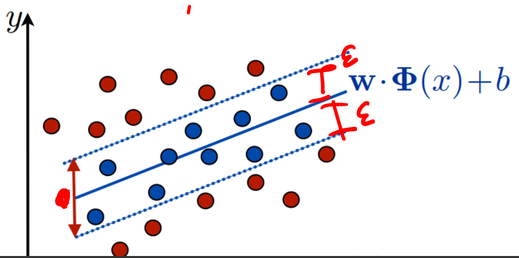
The ϵ**-insensitive loss** is defined as

$\ell(t, y) := \max(0, |y - t| - \epsilon)$

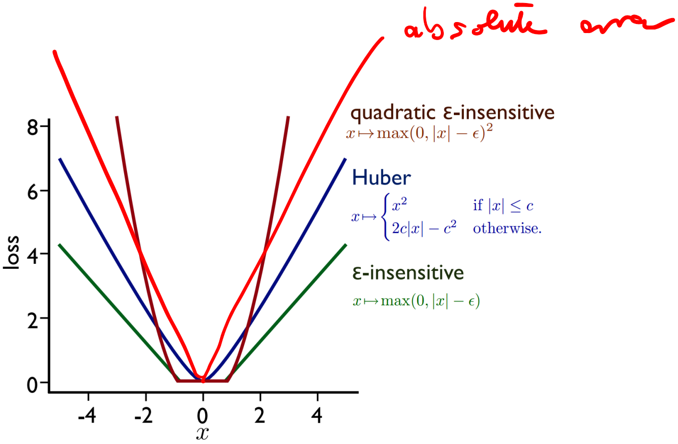
Support Vector Regression (SVR)

$$\mathbf{w}^*_{SVR} := \underset{\mathbf{w}\in\mathcal{H}}{\arg\min} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \max\left(0, |y_i - \langle \mathbf{w}, \phi(\mathbf{x}_i)\rangle| - \epsilon\right)$$

Fit 'tube' with width ϵ to data.



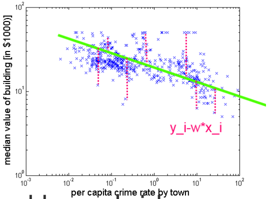
Outlook: More Losses



Conclusion

Regression

- Given $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ and $y_1, \ldots y_n \in \boxed{\mathbb{R}}$, find f such that $f(\mathbf{x}) \approx y$ on new x and y

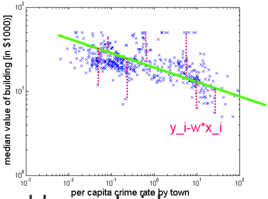


- Important example: ridge regression and kernel ridge regression

$$\min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{2}\|\mathbf{w}\|^2 + C\|\mathbf{y} - X^\top \mathbf{w}\|^2$$

Regression

- Given $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ and $y_1, \ldots y_n \in \boxed{\mathbb{R}}$, find f such that $f(\mathbf{x}) \approx y$ on new x and y



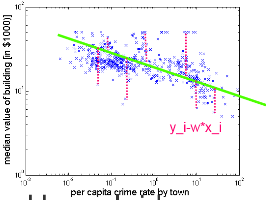
- Important example: ridge regression and kernel ridge regression

$$\min_{\mathbf{w} \in \mathbb{R}^d} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\|\mathbf{y} - X^\top \mathbf{w}\|^2$$

$$= \min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \quad \frac{1}{2}\|X\boldsymbol{\alpha}\|^2 + C\|\mathbf{y} - X^\top X\boldsymbol{\alpha}\|^2$$

Conclusion

Regression



- Given $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ and $y_1, \ldots y_n \in \boxed{\mathbb{R}}$, find f such that $f(\mathbf{x}) \approx y$ on new x and y
- Important example: ridge regression and kernel ridge regression

$$\min_{\mathbf{w} \in \mathbb{R}^d} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\|\mathbf{y} - X^\top \mathbf{w}\|^2$$

$$= \min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \quad \frac{1}{2}\|X\boldsymbol{\alpha}\|^2 + C\|\mathbf{y} - X^\top X\boldsymbol{\alpha}\|^2$$

dxd nxn

- With analytic solutions: $O(d^3)$ $O(n^3)$

$\mathbf{w}_{\mathrm{RR}} = \left(XX^\top + \frac{1}{2C}I\right)^{-1}Xy$, $\boldsymbol{\alpha}^* = \left(K + \frac{1}{2C}I_{n\times n}\right)^{-1}y$

Conclusion (2/2)

LOO~~O~~CV (= amazingly accurate validation) of (K)RR:

- comes (computationally) for free!

LOOOV (= amazingly accurate validation) of (K)RR:

- comes (computationally) for free!

Deep regression:

$$\min_{\mathbf{w},W} \quad \frac{1}{2}\|\mathbf{w}\|^2 \; + \; \frac{1}{2}\sum_{l=1}^{L}\|W_l\|_{\text{Fro}}^2 \; + \; C\sum_{i=1}^{n}(\langle \mathbf{w}, \underset{W}{\phi}(\mathbf{x}_i)\rangle - y_i)^2$$

Conclusion (2/2)

LOOOV (= amazingly accurate validation) of (K)RR:

- comes (computationally) for free!

Deep regression:

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

Unifying view:

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