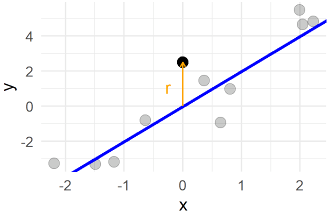
Introduction to Machine Learning

Advanced Risk Minimization
Pseudo-Residuals



Learning goals

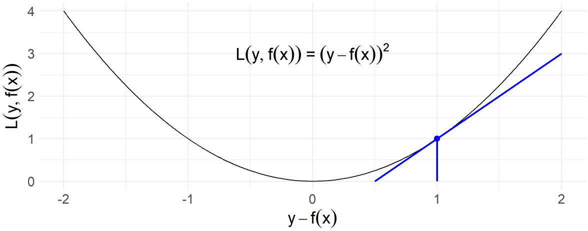
- Concept of pseudo-residuals
- PRs for common losses

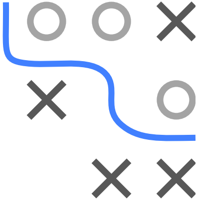
PSEUDO-RESIDUALS

- In regression, residuals are defined as $r(\mathbf{x}) := y - f(\mathbf{x})$
- Generalize concept to **pseudo-residuals**:

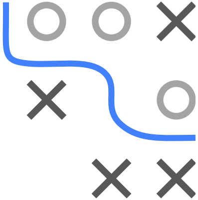
$$\tilde{r}(\mathbf{x}) := -\frac{dL\left(y, f(\mathbf{x})\right)}{df(\mathbf{x})}$$

- If we wiggle $f(\mathbf{x})$, how much does L change?
- Can be used for score-based classifiers and other models
- Note that $\tilde{r}(\mathbf{x})$ depends on y, $f(\mathbf{x})$ and L



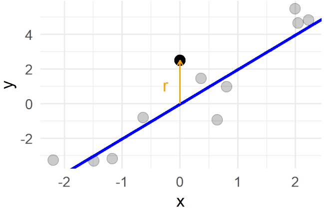


BEST POINT-WISE UPDATE



- Assume we have (partially) fitted a model $f(\mathbf{x})$ to data $\mathcal{D}$
- Assume we could update $f(\mathbf{x})$ point-wise as we like
- Under squared loss, for a fixed $\mathbf{x} \in \mathcal{X}$, the best point-wise update is the direction of the residual $r(\mathbf{x}) = y - f(\mathbf{x})$

$$f(\mathbf{x}) \leftarrow f(\mathbf{x}) + r(\mathbf{x})$$

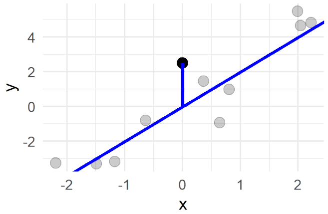


BEST POINT-WISE UPDATE

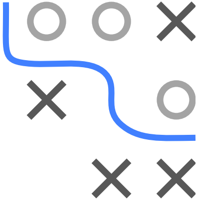
- Assume we have (partially) fitted a model $f(\mathbf{x})$ to data $\mathcal{D}$
- Assume we could update $f(\mathbf{x})$ point-wise as we like
- Under squared loss, for a fixed $\mathbf{x} \in \mathcal{X}$, the best point-wise update is the direction of the residual $r(\mathbf{x}) = y - f(\mathbf{x})$

$$f(\mathbf{x}) \leftarrow f(\mathbf{x}) + r(\mathbf{x})$$

- Point-wise error at this specific $\mathbf{x}$ becomes 0



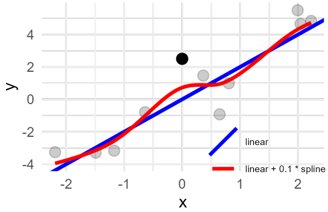
BEST POINT-WISE UPDATE



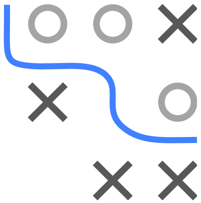
- Assume we have (partially) fitted a model $f(\mathbf{x})$ to data $\mathcal{D}$
- Assume we could update $f(\mathbf{x})$ point-wise as we like
- Under squared loss, for a fixed $\mathbf{x} \in \mathcal{X}$, the best point-wise update is the direction of the residual $r(\mathbf{x}) = y - f(\mathbf{x})$

$$f(\mathbf{x}) \leftarrow f(\mathbf{x}) + r(\mathbf{x})$$

- (In gradient boosting, which we cover later, we don't do point-wise updates but "smoothly distort" f so we generalize)



APPROXIMATE BEST POINT-WISE UPDATE



- Best local change of f at $\mathbf{x}$ to reduce loss most:

$$f(\mathbf{x}) \leftarrow f(\mathbf{x}) - \frac{dL(y, f(\mathbf{x}))}{df(\mathbf{x})}$$

- This is effectively the PR

$$f(\mathbf{x}) \leftarrow f(\mathbf{x}) + \tilde{r}(\mathbf{x})$$

- (Such iterative updates of f like a loss-reducing GD in function space is the major underlying idea of GB)

GD IN ML AND PSEUDO-RESIDUALS

- In GD, we move in the direction of the negative gradient by updating the parameters:

$$\boldsymbol{\theta}^{[t+1]} = \boldsymbol{\theta}^{[t]} - \alpha^{[t]} \cdot \nabla_{\boldsymbol{\theta}} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{[t]}}$$

- Using the chain rule:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \left.\frac{dL(y^{(i)}, f(\mathbf{x}))}{df(\mathbf{x})}\right|_{f=f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)} \cdot \nabla_{\boldsymbol{\theta}} f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right) \\
&= -\sum_{i=1}^{n} \tilde{r}^{(i)} \cdot \nabla_{\boldsymbol{\theta}} f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right).
\end{aligned}$$

- Update is loss-optimal directional change of model output and a loss-independent derivative of $f(\mathbf{x})$
- This is a flexible, nearly loss-independent variant of GD

PSEUDO-RESIDUALS FOR COMMON LOSSES

Loss	Domain of y	Pseudo residual $\tilde{r}$
Squared loss	$y \in \mathbb{R}$	$y - f(\mathbf{x})$
Bernoulli loss	$y \in \{0, 1\}$	$y - s(f(\mathbf{x})) = y - \pi(\mathbf{x})$
Multinomial loss	$y \in \{1, \ldots, g\}$	$\mathbb{1}_{\{y=k\}} - \pi_k(\mathbf{x})$
Exponential loss	$y \in \{-1, 1\}$	$y \exp(-yf(\mathbf{x}))$

NB: $\pi(\mathbf{x}) = s(f(\mathbf{x})) = \frac{\exp(f(\mathbf{x}))}{1+\exp(f(\mathbf{x}))}$ is the (sigmoidal) logistic function, and $\pi_k(\mathbf{x})$ its multi-class generalization, the softmax