Q & A: Regularization

Lecture Series „Machine Learning“

Niels Landwehr

Research Group „Data Science“
Institute of Computer Science
University of Hildesheim

Quiz: Regularization Weight

- Assume a data set $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ with $\mathbf{x} \in \mathbb{R}^M$, $y \in \mathbb{R}$

- We train a linear regression $f_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}$ using squared loss and L2-regularization with a regularization weight $\lambda > 0$:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \sum_{n=1}^{N} (f_{\boldsymbol{\theta}}(\mathbf{x}) - y_i)^2 + \lambda \|\boldsymbol{\theta}\|_2^2$$

- **Question 1**: Which statement typically holds if we decrease the regularization weight λ:
  - The resulting model will have higher squared error on the training set
  - The resulting model will have lower squared error on the training set
  - The resulting model will have the same squared error on the training set
  - Cannot say anything based on the information we have

Quiz: Regularization Weight

- **Solution 1**: The resulting model will typically have lower squared error on the training set
  - Model training trades off between minimizing the squared loss and minimizing the regularization term
  - If we decrease the weight of the regularization term, the optimization focuses more on minimizing the loss, therefore the loss will decrease

Quiz: Regularization Weight

- Assume a data set $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ with $\mathbf{x} \in \mathbb{R}^M$, $y \in \mathbb{R}$
- We train a linear regression $f_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}$ using squared loss and L2-regularization with a regularization weight $\lambda > 0$:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \sum_{n=1}^{N} (f_{\boldsymbol{\theta}}(\mathbf{x}) - y_i)^2 + \lambda \|\boldsymbol{\theta}\|_2^2$$

- **Question 2**: Which statement typically holds if we decrease the regularization weight λ:
  - The resulting model will have higher squared error on the test set
  - The resulting model will have lower squared error on the test set
  - The resulting model will have the same squared error on the test set
  - Cannot say anything based on the information we have

Quiz: Regularization Weight

- **Solution 2**: Cannot say anything based on the information we have
  - Whether the test error decreases or increases depends on where the current regularization weight is in term of making the model underfit or overfit the data
  - If the regularization weight was too high, reducing it reduces test error, if it was too low, reducing it increases test error

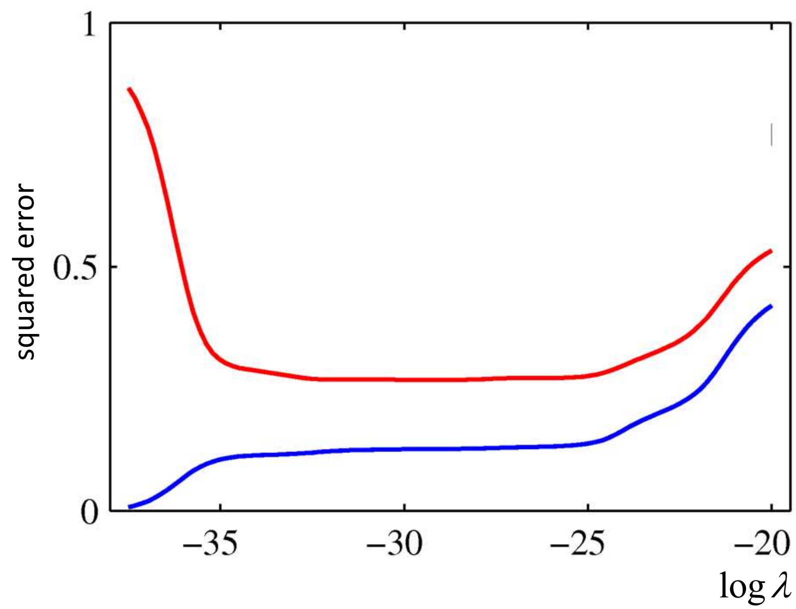

Figure: C. M. Bishop, "Pattern Recognition and Machine Learning", Springer, 2006

Quiz: Variable Selection

- We study the problem of variable selection when learning a boolean function $f:\{0,1\}^2 \to \{0,1\}$ from the following data set $\mathcal{D}$:

	x_1	x_2	y
$(\mathbf{x}_1, y_1)$	0	1	0
$(\mathbf{x}_2, y_2)$	1	0	0
$(\mathbf{x}_3, y_3)$	1	1	1

- We use a finite, discrete space of possible functions $\mathcal{F}$ that consist of all Boolean formulae over the input variables, e.g. $f((x_1,x_2)) = x_1 \wedge x_2$ or $f((x_1,x_2)) = 1$
- We define a learning algorithm $\mathcal{A}(\mathcal{D})$ that given a data set $\mathcal{D}$ returns the model with highest prediction accuracy:

$$\mathcal{A}(\mathcal{D}) = \arg\max_{f \in \mathcal{F}} \frac{1}{3} \sum_{n=1}^{3} I(f(\mathbf{x}_i) = y_i) \qquad \text{where } I(f(\mathbf{x}_i) = y_i) = \begin{cases} 1 : f(\mathbf{x}_i) = y_i \\ 0 : f(\mathbf{x}_i) \neq y_i \end{cases}$$

- As the scoring function we use $\mathcal{S}(f) = \frac{1}{3} \sum_{n=1}^{3} I(f(\mathbf{x}_i) = y_i) - 0.1 \cdot |V|$
- **Question 1**: which subset $V \subseteq \{x_1, x_2\}$ of the original feature set will greedy **backward search** return?

Quiz: Variable Selection

- **Solution 1**: Greedy backward search will return the full feature set $V = \{x_1, x_2\}$

	x_1	x_2	y
$(\mathbf{x}_1, y_1)$	0	1	0
$(\mathbf{x}_2, y_2)$	1	0	0
$(\mathbf{x}_3, y_3)$	1	1	1

- Greedy backward search start with the full feature set $V = \{x_1, x_2\}$
  - The model learned for this feature set is $\mathcal{A}(\pi_V(\mathcal{D})) = x_1 \wedge x_2$
  - The score for this feature set is $\mathcal{S}(\mathcal{A}(\pi_V(\mathcal{D}))) = 0.8$ because this model makes perfect predictions on the training data

- Greedy backward search will then try out the feature sets $V = \{x_1\}$ and $V = \{x_2\}$
  - But those feature sets will result in a lower score of $\mathcal{S}(\mathcal{A}(\pi_V(\mathcal{D}))) \approx 0.56$
  - Therefore the full feature set is kept

Quiz: Variable Selection

- We study the problem of variable selection when learning a boolean function $f:\{0,1\}^2 \rightarrow \{0,1\}$ from the following data set $\mathcal{D}$:

	x_1	x_2	y
$(\mathbf{x}_1, y_1)$	0	1	0
$(\mathbf{x}_2, y_2)$	1	0	0
$(\mathbf{x}_3, y_3)$	1	1	1

- We use a finite, discrete space of possible functions $\mathcal{F}$ that consist of all Boolean formulae over the input variables, e.g. $f((x_1, x_2)) = x_1 \wedge x_2$ or $f((x_1, x_2)) = 1$
- We define a learning algorithm $\mathcal{A}(\mathcal{D})$ that given a data set $\mathcal{D}$ returns the model with highest prediction accuracy:

$$\mathcal{A}(\mathcal{D}) = \arg\max_{f \in \mathcal{F}} \frac{1}{3}\sum_{n=1}^{3} I(f(\mathbf{x}_i) = y_i) \qquad \text{where } I(f(\mathbf{x}_i) = y_i) = \begin{cases} 1 : f(\mathbf{x}_i) = y_i \\ 0 : f(\mathbf{x}_i) \neq y_i \end{cases}$$

- As the scoring function we use $\mathcal{S}(f) = \frac{1}{3}\sum_{n=1}^{3} I(f(\mathbf{x}_i) = y_i) - 0.1 \cdot |V|$
- **Question 2**: which subset $V \subseteq \{x_1, x_2\}$ of the original feature set will greedy **forward search** return?

Quiz: Variable Selection

- **Solution 2**: Greedy forward search will return the empty feature set $V = \emptyset$

	x_1	x_2	y
$(\mathbf{x}_1, y_1)$	0	1	0
$(\mathbf{x}_2, y_2)$	1	0	0
$(\mathbf{x}_3, y_3)$	1	1	1

- Greedy forward search start with the empty feature set $V = \emptyset$
  - The model learned for this feature set is $\mathcal{A}(\pi_V(\mathcal{D})) = 0$
  - The score for this feature set is $\mathcal{S}(\mathcal{A}(\pi_V(\mathcal{D}))) \approx 0.66$ because this model makes perfect predictions on the training data
- Greedy forward search will then try out the feature sets $V = \{x_1\}$ and $V = \{x_2\}$
  - But adding a single feature will result in a lower score: $\mathcal{S}(\mathcal{A}(\pi_V(\mathcal{D}))) \approx 0.56$
  - Therefore the empty feature set is kept