02 - Linear Models

Numerical Methods for Deep Learning

January 7, 2018

Classification and least-squares regression

Give, examples

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1^\top \\ \mathbf{y}_2^\top \\ \vdots \\ \mathbf{y}_n^\top \end{pmatrix} \in \mathbb{R}^{n \times n_f}$$

and labels

$$\mathbf{C} = \begin{pmatrix} \mathbf{c}_1^\top \\ \mathbf{c}_2^\top \\ \vdots \\ \mathbf{c}_n^\top \end{pmatrix} \in \mathbb{R}^{n \times n_c}$$

Goal: Find a classification/prediction function $f(\cdot, \boldsymbol{\theta})$, i.e.,

$$f(\mathbf{y}_j, \boldsymbol{\theta}) \approx \mathbf{c}_j, \quad j = 1, \ldots, n.$$

Regression and least-squares

Simplest option, a linear model

$$\mathbf{YW} + \mathbf{1b}^\top = \mathbf{C}$$

- $\mathbf{W} \in \mathbb{R}^{n_f \times n_c}$ are *weights*
- $\mathbf{b} \in \mathbb{R}^{n_c}$ are *biases*
- $\mathbf{1} \in \mathbb{R}^{n}$ is a vector of ones

Equivalent notation:

$$\begin{pmatrix} \mathbf{Y} & \mathbf{1} \end{pmatrix} \begin{pmatrix} \mathbf{W} \\ \mathbf{b}^\top \end{pmatrix} = \mathbf{C}$$

Problem may not have a solution, or may have infinite solutions (when?). Solve through optimization

$$\min_{\mathbf{W}} \frac{1}{2} \|\mathbf{YW} - \mathbf{C}\|_F^2$$

(Frobenius norm: $\|\mathbf{A}\|_F^2 = \mathrm{trace}(\mathbf{A}^\top \mathbf{A}) = \sum_{i,j} \mathbf{A}_{i,j}^2$.)

Regression and least-squares

To minimize a function need to differentiate and equate to 0

$$\frac{\partial\left(\frac{1}{2}\|\mathbf{YW}-\mathbf{C}\|_F^2\right)}{\partial\mathbf{W}} = 0$$

Computing the derivatives in three steps

1. $$\frac{\partial\left(\frac{1}{2}\|\mathbf{R}\|_F^2\right)}{\partial\mathbf{R}} = ???$$
2. $$\frac{\partial\left(\mathbf{YW}\right)}{\partial\mathbf{W}} = ???$$
3. Use chain rule

Regression and least-squares

Putting it all together gives

$$\frac{\partial\left(\frac{1}{2}\|\mathbf{YW}-\mathbf{C}\|_F^2\right)}{\partial\mathbf{W}}=\mathbf{Y}^\top(\mathbf{YW}-\mathbf{C})=0$$

Reorganizing obtain the **normal equations**

$$\mathbf{W}=(\mathbf{Y}^\top\mathbf{Y})^{-1}\mathbf{Y}^\top\mathbf{C}.$$

Assume that $\mathbf{Y}^\top\mathbf{Y}$ is invertible

- Sufficient amount of data
- Data is linearly independent

Coding: Least-Squares

Outline

- dataset: MNIST / CIFAR10 /Segmentation
- Least-squares via normal equations

Ill-posedness and regularization - I

If the data is linearly dependent or close to be linearly dependent, least-squares problem gives no good solution. Understanding can be gained by the Singular Value Decomposition (SVD)

$$\mathbf{Y} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$$

where $\mathbf{U} \in \mathbb{R}^{n \times n_f}, \mathbf{V} \in \mathbb{R}^{n_f \times n_f}$ satisfy

$$\mathbf{U}^\top\mathbf{U} = \mathbf{I}, \quad \text{and} \quad \mathbf{V}^\top\mathbf{V} = \mathbf{I}$$

And $\boldsymbol{\Sigma}$ contains the singular values along diagonal

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_{n_f} \end{pmatrix}$$

Ill-posedness and regularization - I

Important is the effective rank: If $\sigma_j \ll \sigma_1$ for all $j \geq k$, then the effective rank of the problem is k.

If $k < n_f$, the least squares problem is ill-posed, i.e., solution does not exist or is unstable.
Small perturbations in $\mathbf{C}$ or $\mathbf{Y}$ yield large perturbations in $\mathbf{W}$

Solve regularized problem: For $\alpha > 0$

$$\min_{\mathbf{W}} \frac{1}{2}\|\mathbf{YW} - \mathbf{C}\|_F^2 + \frac{\alpha}{2}\|\mathbf{W}\|_F^2$$

Class assignment - solve the optimization problem

Ill-posedness and regularization - I

Important is the effective rank: If $\sigma_j \ll \sigma_1$ for all $j \geq k$, then the effective rank of the problem is k.

If $k < n_f$, the least squares problem is ill-posed, i.e., solution does not exist or is unstable.
Small perturbations in $\mathbf{C}$ or $\mathbf{Y}$ yield large perturbations in $\mathbf{W}$

Solve regularized problem: For $\alpha > 0$

$$\min_{\mathbf{W}} \frac{1}{2}\|\mathbf{YW} - \mathbf{C}\|_F^2 + \frac{\alpha}{2}\|\mathbf{W}\|_F^2$$

Class assignment - solve the optimization problem

$$\mathbf{W} = (\mathbf{Y}^\top \mathbf{Y} + \alpha \mathbf{I})^{-1} \mathbf{Y}^\top \mathbf{C}$$

The bias variance decomposition

Regularization and α

Assume $\mathbf{C} = \mathbf{Y}\mathbf{W}_{\text{true}} + \epsilon$
Then setting $\mathbf{Y}^{\dagger}_{\alpha} = (\mathbf{Y}^{\top}\mathbf{Y} + \alpha\mathbf{I})^{-1}$

$$
\begin{aligned}
\mathbf{W} - \mathbf{W}_{\text{true}} &= \mathbf{Y}^{\dagger}_{\alpha}\mathbf{Y}^{\top}\mathbf{C} - \mathbf{W}_{\text{true}} \\
&= \left(\mathbf{Y}^{\dagger}_{\alpha}\mathbf{Y}^{\top}\mathbf{Y} - \mathbf{I}\right)\mathbf{W}_{\text{true}} + \mathbf{Y}^{\dagger}_{\alpha}\mathbf{Y}^{\top}\epsilon \\
&= -\alpha\mathbf{Y}^{\dagger}_{\alpha}\mathbf{W}_{\text{true}} + \mathbf{Y}^{\dagger}_{\alpha}\mathbf{Y}^{\top}\epsilon
\end{aligned}
$$

The bias variance decomposition

Depends on a random variable ϵ - take expectation

$$\mathbb{E}\|\mathbf{W} - \mathbf{W}_{\text{true}}\|^2 = \mathbb{E}\|\mathbf{Y}_\alpha^\dagger \mathbf{Y}^\top \epsilon - \alpha \mathbf{Y}_\alpha^\dagger \mathbf{W}_{\text{true}}\|^2 =$$

$$\overbrace{\alpha^2 \|\mathbf{Y}_\alpha^\dagger \mathbf{W}_{\text{true}}\|^2}^{\|\text{bias}\|^2} + \overbrace{\sigma^2 \text{trace}\left(\mathbf{Y}\mathbf{Y}_\alpha^{\dagger^T} \mathbf{Y}_\alpha^\dagger \mathbf{Y}^\top\right)}^{\text{variance}}$$

Point to take home

No such thing as exact recovery!