Optimization in Data Analysis—project's topics

Consider the binary classification problem in which we have 2 categories (classes A and B) denoted by variables z (objects are described by the feature vector $y \in \mathcal{R}^n$.):

$$z = \begin{cases} 1 & \text{if } y \text{ is in category } A \\ -1 & \text{if } y \text{ is in category } B \end{cases}.$$

Assume that (y_i, z_i), $i = 1, \ldots, N$ are empirical data related to the feature vector and the variable denoting the class.

Object's category is decided on the basis of the decision function $h(x, y)$ which parameters x are found by solving the optimization problem:

$$\min_x \left[f(x) \equiv r(x) + C \sum_{i=1}^{N} \xi(x; y_i, z_i) \right], \tag{1}$$

where

- r is the regularization component which guarantees the 'proper profile' of x,
- C is a constant (hyperparameter) which provides 'good' balance between r and the loss function.

Examples of loss functions ξ (notice that we assume that $z \in \{1, -1\}$)

- Support Vector Machine (SVM) with loss function $L1$

$$\xi_1(x; y, z) \equiv \max(0, 1 - zx^T y) \tag{2}$$

- SVM with loss function $L2$

$$\xi_2(x; y, z) \equiv \left[\max(0, 1 - zx^T y)\right]^2 \tag{3}$$

- Logistic regression (LR)

$$\xi_{LR}(x; y, z) \equiv \log(1 + e^{-zx^T y}) \tag{4}$$

Most common regularization components:

- $L2$ regularization

$$r_{L2}(x) \equiv \|x\|_2^2 = \sum_{i=1}^{n} x_i^2 \tag{5}$$

- $L1$ regularization

$$r_{L1}(x) \equiv \|x\|_1 = \sum_{i=1}^{n} |x_i| \tag{6}$$

Project's related to a linear classification

C1 (gr. 1) Binary classification problem with the loss function ξ_1 and regularization r_{L1}—the problem solved by interior point methods:

$$\min_x [C\xi_1 + r_{L1}]$$

C2 (gr. 1) Binary classification problem with the loss function ξ_2 and the regularization r_{L1}—the problem solved by the method described in [K.-W. Chang, C.-J. Hsieh, and C.-J. Lin. Coordinate Descent Method for Large-scale L2-loss Linear SVM. Journal of Machine Learning Research 9(2008), 1369-1398]:

$$\min_x [C\xi_2 + r_{L1}]$$

C3 (gr. 1) Binary classification problem with the loss function ξ_1 and regularization r_{L2}:

$$\min_x [C\xi_1 + r_{L2}]$$

The problem solved by the method described in [V. Franc, S. Sonnenburg. Optimized cutting plane algorithm for support vector machines. In Proceedings of the 25th international conference on Machine learning (ICML '08). Association for Computing Machinery, 2008, New York, NY, USA, pp. 320–327].

C4 (gr. 1) Binary classification problem with the loss function ξ_2 and regularization r_{L2}:

$$\min_x [C\xi_2 + r_{L2}]$$

C5 (gr. 1) Binary classification problem with the loss function ξ_{LR} and regularization r_{L1}:

$$\min_x [C\xi_{LR} + r_{L1}]$$

C6 (gr. 2) Binary classification problem with the loss function ξ_{LR} and regularization r_{L2}:

$$\min_x [C\xi_{LR} + r_{L2}]$$

C7 (gr. 2) Binary classification problem with the loss function ξ_{LR} and regularization r_{L1}:

$$\min_x [C\xi_{LR} + r_{L1}]$$

The problem transformed to the problem with a smooth function and box constraints

C8 (gr. 2) Binary classification problem with the loss function ξ_1 and regularization r_{L1}:

$$\min_x [C\xi_1 + r_{L1}]$$

The problem solved cutting plane methods ([C. Teo, et al., Bundle methods for regularized risk minimization. Journal of Machine Learning Research, Vol. 11, pp. 311—365, 2010])

C9 (gr. 2) Binary classification problem with the loss function ξ_1 and regularization r_{L2}:

$$\min_x [C\xi_1 + r_{L2}]$$

The problem solved by 'coordinate descent' method ([K–W. Chang, et al, Coordinate descent method for large-scale l2-loss linear support vector machines. Journal of Machine Learning Research, Vol. 9, pp. 1369-–1398, 2008]).

C10 (gr. 1) Binary classification problem with the loss function ξ_1 and regularization r_{L2}:

$$\min_x [C\xi_1 + r_{L2}]$$

The problem solved by transforming to the dual problem and then using the method [C–Jui Hsieh, et al, A dual coordinate descent method for large-scale linear svm, Proceedings of the 25th international conference on Machine learning, pp. 408-–415, 2008]

C11 (gr. 2) Binary classification problem with the loss function ξ_1 and regularization r_{L2}:

$$\min_x [C\xi_1 + r_{L2}]$$

The problem solved by transforming to the dual problem and then using solvers for quadratic problems.

Consider linear regression model

$$y = X\beta + e, \tag{7}$$

where $y \in \mathcal{R}^N$ is a vector of observations (empirical data) of endogenous variables, $X \in \mathcal{R}^{N\times p}$ is a matrix of observations of exogenous variables, $\beta \in \mathcal{R}^p$ is a vector of model's parameters and $e \in \mathcal{R}^N$ is model's error.

We build a linear regression model by solving the optimization problem

$$\min_\beta \frac{1}{2}\|y - X\beta\|_2^2, \text{ subject to the constraints } \|\beta\|_0 \le k, \tag{8}$$

where k is a given natural number not greater than p.

Here $\|\beta\|_0$ denotes the number of nonzero elements of the vector β thus solving the problem (8) gives us the best model's parameters in the sense of the least–squares error under additional assumption that the number of the model's parameters cannot be greater than k:

$$\|\beta\|_0 = \sum_{i=1}^{p} \mathcal{I}(\beta_i \neq 0),$$

where $\mathcal{I}(\cdot)$ is the indicator function.

Notice that we aim at building regression models with as few as possible parameters (each exogenous variable in the model requires its forecast).

Approximate solution of the problem (8) can be achieved by solving the problem

$$\min_\beta \left[\frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1\right], \tag{9}$$

where λ is a penalty parameter.

This is convex, nondifferentiable optimization problem.

Project's topics related to regressions models:

R1 (gr. 1) The problem (9) solved by own implementations of 'coordinate descent' methods based on [Hao-Jun Michael Shi, et al, A Primer on Coordinate Descent Algorithms, arXiv:1610.00040, 2017].

R2 (gr. 2) The problem (9) solved by the Nesterov's method ([Nesterov, Mathematical Programming, Vol. 140, 2013, pp. 126–161])

R3 (gr. 2) The problem (9) solved by 'pathwise coordinate' method ([Friedman et al, The Annals of Applied Statistics, Vol. 1, 2007, pp. 302–322])

R4 (gr. 2) The problem

$$\min_{\beta} \left[\frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right], \tag{10}$$

solved by the method described in [Yuan, Lin, Journal of Royal Statistical Society B, Vol. 68, 2006, pp. 46–67].

R5 (gr. 2) The problem (9) solved by the method described in [Efron, et al, The Annals of Statistics, Vol. 32 2004, pp. 407–499]

R6 (gr. 1) The problem (9) solved by the method described in [Efron, et al, The Annals of Statistics, Vol. 32 2004, pp. 407–499] and with the help of ROI (R Optimization Infrastructure)

R7 (gr. 1) The problem (8) solved by mixed integer programming method ([Bertsimas, et al, The Annals of Statistics, Vol. 44, 2016, pp. 813–852]), the case $n > p$

R8 (gr. 2) The problem (8) solved by mixed integer programming method ([Bertsimas, et al, The Annals of Statistics, Vol. 44, 2016, pp. 813–852]), the case $p \gg n$

R9 (gr. 1) The problem (8) solved by first order methods ([Bertsimas, et al, The Annals of Statistics, Vol. 44, 2016, pp. 813–852])

Project's topics related to use optimization in deep learning:

DL1 (gr. 1) Learning neural networks using Stochastic Gradient Descent with Nesterov's momentum ([Sutskever, I., Martens, J., Dahl, G., and Hinton, G. (2013). On the importance of initialization and momentum in deep learning. In ICML. 296, 401, 408], [Nesterov, 1983, 2004, 2018]). Comparison with ADAM's version of SGD.

DL2 (gr. 2) Learning neural networks using Stochastic Gradient Descent with tuning hyperpaarameters—AdaGrad method ([Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research. 303]). Comparison with ADAM's version of SGD.

DL3 (gr. 1) Learning neural networks using Stochastic Gradient Descent with tuning hyperpaarameters—RMSProp method ([Hinton, Lectures, 2012]). Comparison with ADAM's version of SGD.

DL4 (gr. 2) Learning neural networks using Stochastic Gradient Descent with tuning hyperpaarameters—RMSProp method ([Hinton, Lectures, 2012]). Comparison with ADAM's version of SGD. Numerical experiments performed on datasets related to fine tuning of pretrained BioBERT model aimed at Named Entity Recognition task ([J. Lee, et al, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics, 2019]). Project for 3 students.

DL5 (gr. 1) Learning neural networks using Stochastic Gradient Descent with tuning hyperpaarameters—RMSProp method ([Hinton, Lectures, 2012]). Comparison with ADAM's version of SGD. Numerical experiments performed on datasets related to fine tuning of pretrained BioBERT model aimed at Relation Extraction task ([J. Lee, et al, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics, 2019]). Project for 3 students.

DL6 (gr. 2) Learning neural networks using Stochastic Gradient Descent with tuning hyperpaarameters—RMSProp method ([Hinton, Lectures, 2012]). Comparison with ADAM's version of SGD. Numerical experiments performed on datasets related to fine tuning of pretrained BioBERT model aimed at Question Answering task ([J. Lee, et al, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics, 2019]). Project for 3 students.

DL7 (gr. 1) Learning neural networks using Stochastic Gradient Descent with tuning hyperpaarameters—AdaGrad method ([J. Duchi, et al, Adaptive subgradient methods for online learning and stochastic optimization. J Mach Learn Res. 2011;12, 2121–59]). Comparison with ADAM's version of SGD. Numerical experiments performed on datasets related to fine tuning of pretrained CollaboNet model aimed at Named Entity Recognition task ([W. Yoon, et al, CollaboNet: collaboration of deep neural networks for biomedical named entity recognition, BMC Bioinformatics 2019, 20(Suppl 10):249]). Project for 4 students.