Overview

The below assignment consists of theoretical and practical parts. Please prepare the solutions for the theoretical part as PDF scans of your derivations on paper (alternatively you can use any software, e.g. LaTeX, to generate PDF with formulas). For the practical part, please prepare the solution in jupyter notebook format. The solutions should be clear and easy to follow.

Please send your PDF and .ipynb files to amaevskij@hse.ru under the following subject:
MLDM-2019-HW1-<YOUR LASTNAME>.

The deadline is: **23:59, 24.10.2019**.

Note the total number of points you get will be clipped at 1 from above:

$$\text{total score} = \min\left[1,\ \sum_i \text{score for exercise } \#i\right].$$

1 Theory

1.1 MSE and MAE (0.1 points)

Given a set of points $y_i \in \mathbb{R}$, $i = 1, ..., n$, find $\hat{y}_{\text{MSE}}$ and $\hat{y}_{\text{MAE}}$ minimizing MSE and MAE losses, respectively:

$$\text{MSE}(\hat{y}_{\text{MSE}}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{\text{MSE}})^2,$$

$$\text{MAE}(\hat{y}_{\text{MAE}}) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_{\text{MAE}}|.$$

1.2 Solution for the Ridge Regression (0.15 points)

Given a design matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ and a vector of targets $\mathbf{y} \in \mathbb{R}^n$, derive the analytical solution for the Ridge Regression problem, i.e. solve:

$$\|\mathbf{Xw} - \mathbf{y}\|^2 + \lambda\|\mathbf{w}\|^2 \underset{\mathbf{w}}{\longrightarrow} \min$$

for the model parameters $\mathbf{w} \in \mathbb{R}^d$.

1.3 Likelihood vs cross-entropy (0.1 points)

Show that for a given parametric probability distribution $\hat{P}(x\,|\,\theta)$ maximum likelihood estimation of θ is equivalent to minimizing cross-entropy between the empirical probability distribution $P_{\text{data}}(x)$ and $\hat{P}(x\,|\,\theta)$.

1.4 Optimism of the training MSE (0.25 points)

Show that for a training set of size N the average optimism ω of MSE is given by:

$$\omega = \frac{2}{N}\sum_{i=1}^{N}\underset{\tau_y}{\text{cov}}(\hat{y}_i, y_i),$$

where $y_i \in \mathbb{R}$ and $\hat{y}_i \in \mathbb{R}$ are the true and predicted target values for the object i, respectively, and subscript τ_y means that expectations in the covariance are taken only with respect to the targets y_i, keeping the features x_i fixed. *See [1], chapter 7, exercise 7.4 for a hint.*

1.5 Linear Regression optimism (0.15 points)

Show that for linear regression the average optimism ω is reduced to:

$$\omega = \frac{2d}{N}\sigma^2,$$

where d is the dimensionality of the data (number of features), N is the number of training objects and σ is the standard deviation of the noize in the data.

2 Practice

2.1 AdaBoost (0.25 points)

For the task of odd vs even number classification on the MNIST dataset write your own implementation of the AdaBoost algorithm. Use `sklearn.linear_model.LogisticRegression` trained on a random subset of features as the base learner.

2.2 Gradient boosting (0.3 points + 0.1 bonus)

For the task of odd vs even number classification on the MNIST dataset write your own implementation of gradient boosting with cross-entropy loss. Use `sklearn.linear_model.LogisticRegression` trained on a random subset of features as the base learner. Bonus points for using quadratic loss approximation.

2.3 Bias-variance decomposition (0.2 points)

For a classification algorithm of your choice study the dependence of the bias, variance and expected prediction error (0-1 loss) as the function of model complexity (plot these dependencies on the same graph). For this task, generate some toy data.

Getting the MNIST dataset

To get the MNIST data you can use:

```
from tensorflow.keras.datasets import mnist
(x_train, y_train), (x_test, y_test) = mnist.load_data()
```

References

[1] HASTIE, T., TIBSHIRANI, R., AND FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, https://web.stanford.edu/~hastie/ElemStatLearn/, 2009.