Machine Learning 2
Exercise Sheet 6

Prof. Dr. Dr. Lars Schmidt-Thieme, Brad Baker
Information Systems and Machine Learning Lab
University of Hildesheim

May 24th, 2016

Submission until May 30th, 8:00 AM by learnweb.
Please put your name in all filenames.
Non-pdf submissions for non-programming exercises will not be graded.

Exercise 11: Model Stacking (10 points)

a) (2 points) Suppose we have learned five different classification models, and have evaluated the errors of each of these models. The output of my error evaluation provides a vector for each model which has N_{VALID} many entries, where an entry of -1 indicates an error, and an entry of 1 indicates a success.

We want to choose an ensemble of **three** models which is as effective as possible, and are am using a voting scheme to select the final class label. Given the following errors on validation data of size 6, which models should be incorporated into the ensemble?

$$\begin{aligned}
\hat{y}_1 &: (-1, -1, 1, -1, 1, -1) \\
\hat{y}_2 &: (1, 1, 1, -1, 1, -1) \\
\hat{y}_3 &: (1, -1, -1, 1, 1, 1) \\
\hat{y}_4 &: (1, 1, 1, -1, -1, -1) \\
\hat{y}_5 &: (-1, 1, 1, 1, -1, 1)
\end{aligned}$$

b) (6 points) Suppose I have learned 3 different regression models, $\hat{y}_c$ which give me the following output for 4 different training samples, and which have a ground-truth y:

$$\begin{aligned}
\hat{y}_1 &: (4, 4, 1, 4) \\
\hat{y}_2 &: (3, 1, 2, 2) \\
\hat{y}_3 &: (5, 3, 2, 3) \\
y &: (3, 4, 2, 2).
\end{aligned}$$

- Compute the squared loss of an ensemble model which initially computes $\hat{y}$ using model averaging.
- Then, set up a linear-stacking learning algorithm to learn the parameters α_c for each of the three models. Initialize the parameters as
$$\alpha : (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}).$$
Compute one iteration of your learning algorithm and compare the change in the loss.

c) (2 points) Suppose I have learned an ensemble of models, have analyzed the covariance between the predictions made by all the models, and have discovered 1) that the outputs are identically gaussian distributed, and 2) that when I consider two models $\hat{y}_i, \hat{y}_j$ in my ensemble:

$$cov[\hat{y}|(\hat{y}_i, \hat{y}_j)] = \kappa(\hat{y}_i, \hat{y}_j)$$

where κ is the squared exponential kernel. What kind of secondary model would make a sensible choice for stacking this ensemble? What if κ is a linear kernel?

Exercise 12: Bootstrap and Random Forests

a) (1 point) Suppose we have a training data set, X, with ground-truth y:

$$X = \begin{pmatrix} 13 & 12 \\ 15 & 20 \\ 16 & 41 \\ 32 & 3 \\ 30 & 33 \end{pmatrix}, y = \begin{pmatrix} 7 \\ 9 \\ 8 \\ 2 \\ 9 \end{pmatrix}$$

Give an example of two data sets which can be sampled from this data using bootstrap sampling. You can either do the sampling yourself using a pseudo-random process, or simply provide examples which illustrate the kind of behavior one often sees with bootstrapping.

b) (6 points) Suppose we have learned a Random Forest over M many independent and identically distributed (i.i.d) random variables, each of which has a variance of σ^2. Since these variables are statistically independent, the average variance over the entire model is equal to $\frac{\sigma^2}{M}$.

If the variables are **identically, but not independently** distributed, and have a pairwise correlation of ρ, show that the average variance can be computed as

$$\rho\sigma^2 + (1-\rho)\frac{\sigma^2}{M}.$$

How does this derived variance illustrate the idea behind Random Forests - i.e. the benefits of learning the averaged model over a single decision tree?

c) (3 points) One of the general downsides of Random Forests and other so-called 'Black Box' models is that the resulting models are not always directly interpretable. One method often used to bridge the interpretability-gap is the "Variable Importance" of the variables within the Random Forest model.

- Read Louppe, et. al 2014 for a great overview of the role of this measure in Random Forests.
- In your own words, describe the variable importance measure and how it helps describe Random Forest models.

No assigned bonus this week