Statistical Learning, Fall 2024-5

Practice final exam No. 1

Note: Some of the problems are based on problems from the homework assignments, adjusted to the level that the final is expected to have. The real final is unlikely to have problems from the homework.

1. (30 points) Assume we have a univariate model with one x variable and no intercept. We fit **constrained** ridge regression and lasso with a given constraint s on the norm (ℓ_2 norm squared for ridge, ℓ_1 norm for lasso). Now we add a second identical variable $x^* = x$ and refit the models with the same constraint. What happens to the coefficients $\hat{\beta}$ of both models? How does the two-dimensional solution to the new problem relate to the one-dimensional solution to the old one in each case? Is it unique? Assume the constraint is much smaller than the norm of the least squares solution, so it is tight.
   Note that this problem deals with the constrained form of the regularized problems, not penalized. **Hint:** The behavior of ridge and lasso under this scenario is quite different. Since both predictors x, x^* are identical, a coefficient can be divided between them in different ways which give the same fit. Consider what different divisions do to the norm of the coefficient vector in each case, and use that to infer the optimal solution.

2. (20 points) Assume as usual $y = f(x) + \epsilon$, with $\epsilon \sim (0, \sigma^2)$ independent of x. Prove that the effective degrees of freedom (that is, optimism divided by $2\sigma^2$) of k-NN with N observations is N/k.

3. (25 points) Short intuition problems: Answer and explain briefly (there may be more than one right answer). If you need additional assumptions to reach your conclusion, specify them.

   (a) If I have a very large amount of data in a reasonably low dimension (very large n, small p), which regression method would be most likely to minimize my model's EPE?

      i. Linear regression
      ii. 1-NN (i.e., nearest neighbor with 1 neighbor)
      iii. k-NN with $k = \log(n)$
      iv. k-NN with $k = \sqrt{n}$

   (b) If I believe that only a small number of my variables are important, which one (or more) of these four regularization approaches should I use?

      i. Ridge
      ii. Lasso
      iii. Variable selection
      iv. PCA regression

   (c) We are given a problem (like say our Netflix example) with training set of size 8000 and test set of size 2000. Ruth and Naomi each build a regression tree on this training data, but each uses a different half of the variables/columns (say, Ruth uses the even ones and Naomi the odd ones). They both use all 8000 observations/rows. Obviously the trees look completely different, as the two trees never use the same variables. When applying their models to the test set, they are surprised to find out that the predictions are very similar, say with correlation 0.98 between them. Ruth says "Uh-oh, this makes no sense – we must have a bug", and Naomi replies:"Actually, it means that two different models reach the same conclusions, so our models must be very accurate!". Which one of them is correct and why? Note they may both be correct, or both wrong.

(d) Tom and Jerry are building a random forest on some training data, they obviously want it to predict well on some test data for their competition. Tom says: "Let's hold out some validation data, so we can know how many trees to include in our random forest", and Jerry replies: "Actually, let's not hold out any data, and run it until our it's time to submit our entry for the competition." Which one of them is correct and why? Note they may both be correct, or both wrong.

(e) (*+5 points) Jerry says to Tom: "Actually, I can validate the random forest model without holding out any data from the training." Tom replies: "You can't do it properly, though." Explain clearly and accurately in which sense each of them is correct.
Note: Only short and accurate explanations will receive credit.

4. (25 points) In class we discussed gradient boosting as fitting trees to loss function derivatives. For squared error loss, this amounted to fitting the residual each time. Now assume instead of fitting the residuals $r_i = y_i - \hat{y}_i$ in each boosting iteration, we consider two other choices for fitting:

$$\tilde{r}_i = \begin{cases} r_i \text{ if } |r_i| < c \\ 0 \text{ if } |r_i| \geq c \end{cases} .$$

$$\breve{r}_i = \begin{cases} r_i \text{ if } |r_i| < c \\ c \text{ if } r_i \geq c \\ -c \text{ if } r_i \leq -c \end{cases} .$$

For each of these two cases, analyze the loss function implied by this gradient boosting algorithm in terms of:

- Is it continuous? Differentiable? Convex?
- Does it make sense as a loss function for regression (think about issues like where it attains its minimum and any other relevant aspects)?