Statistical Learning - Milan, Fall2019

# Homework problems 5+6

5 **The effect of identical predictors in Ridge and Lasso (3.28,3.29 in ESL)**
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 ($\ell_2$ norm squared for ridge, $\ell_1$ 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.
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. You can also simulate to gain intuition.

6 **Which properties of Lasso path generalize to other loss functions?**
Recall we showed the optimality conditions for a Lasso solution:

$$\hat{\beta}(\lambda)_k \neq 0 \quad \Rightarrow \quad X_k^T(Y - X\hat{\beta}(\lambda)) = \frac{\lambda}{2}\text{sgn}(\hat{\beta}(\lambda)_k) \tag{1}$$

$$\hat{\beta}(\lambda)_k = 0 \quad \Leftarrow \quad |X_k^T(Y - X\hat{\beta}(\lambda))| < \frac{\lambda}{2} \tag{2}$$

$$\forall k \qquad |X_k^T(Y - X\hat{\beta}(\lambda))| \leq \frac{\lambda}{2}, \tag{3}$$

where as we noted in class,

$$X_k^T(Y - X\hat{\beta}(\lambda)) = -\frac{\partial RSS(\beta)}{\partial \beta_k}|_{\beta=\hat{\beta}(\lambda)}$$

is the derivative of the loss function.

We noted in class the following properties of the set of solutions $\{\hat{\beta}(\lambda) : 0 \leq \lambda \leq \infty\}$:

i All the variables in the solution are "highly correlated" with the current residual from (1) above, and all the variables with zero coefficients are "less correlated" with the current residual from (2,3) above.

ii The solution path $\{\hat{\beta}(\lambda) : 0 \leq \lambda \leq \infty\}$ as a function of $\lambda$ can be described by a collection of "breakpoints" $\infty > \lambda_1 > \lambda_2 > ... > \lambda_K > 0$ such that the set $\mathcal{A}_k$ of active variables with non-zero coefficients is fixed for all solutions $\hat{\beta}(\lambda)$ with $\lambda_k \geq \lambda \geq \lambda_{k+1}$.

iii $\hat{\beta}(\lambda)$ is a piecewise linear function, in other words, for $\lambda$ in this range we have:

$$\hat{\beta}(\lambda) = \hat{\beta}(\lambda_k) + v_k(\lambda_k - \lambda),$$

for a vector $v_k$ we explicitly derived in class.

Assume now that we want to build a different type of model with a different convex and infinitely differentiable loss function, say a logistic regression model for a binary classification task, and add lasso penalty to that:

$$\hat{\beta}(\lambda) = \arg\min_{\beta} \sum_{i=1}^{n} \log\left\{1 + \exp\{-y_i x_i^T \beta\}\right\} + \lambda \|\beta\|_1.$$

We would like to investigate which of the properties above still holds for the solution of this problem.

(a) Using simple arguments about derivatives and sub-derivatives as we used in class for the quadratic loss case, argue that that three conditions like (1)-(3) can be written for this case too, with the appropriate derivative replacing the empirical correlation. Derive these expressions explicitly for the logistic case.

(b) Explain clearly why this implies that properties (i), (ii) still hold (for (ii), you may find the continuity of the derivative useful).

(c) Does the piecewise linearity still hold? A clear intuitive explanation is sufficient here.
**Hint:** Consider how we obtained the linearity for squared loss in $\triangle\lambda$ in class by decomposing the correlation vector $X^T(Y - X\beta) = X^T Y - X^T X\beta$.