Yale University
Department of Statistics and Data Science

STATISTICS 365/565

Issued: 03/29/2021 **Due:** 04/08/2021

Notation: $[k] = \{1, 2, \ldots, k\}$. For a matrix $A \in \mathbb{R}^{m \times n}$ we will let $A_{(i,\cdot)}$ denote the i^{th} row and $A_{(\cdot,j)}$ denote the j^{th} column. **Both will be treated as column vectors.**

1 Deriving adaboost from forward stagewise classification

In lecture we saw that boosting is a form of stagewise additive model. The idea is that at time t we have access to a *weak* learner that can do an ok job optimizing an objective over our data, but not necessarily a great job. We can "boost" those weak learners to create a strong learner. Typically the weak learner is taken to be a shallow (depth 7) decision tree. One of the first methods that demonstrated the possibility of this idea was adaboost. Adaboost is a special case of boosting where we focus on classification with the exponential loss as our surrogate to the Hamming loss. We will derive it as a forward stage-wise additive model.

Please see Section 10.1 of Elements of Statistical Learning for a description of the adaboost algorithm.

a) Suppose that at round m we have the function $G(x) = \sum_{j=1}^{m-1} \alpha_j G_j(x)$. We wish to update with a new additive function $G_m(x)$. We do so by computing

$$\alpha_m, G_m = \arg \min_{\alpha, F \in \mathcal{F}} \sum_{i=1}^{n} \exp(-y_i(G(x_i) + \alpha F(x_i)))$$

where $\mathcal{F}$ is our function class of weak-learners and concretely it is a set of functions that map $\mathcal{X} \to \{-1, +1\}$.

Let $w_i = \exp(-y_i G(x_i))$. Prove that regardless of the choice of α_m, we have

$$G_m = \arg \min_{F \in \mathcal{F}} \sum_{i=1}^{n} w_i \mathbb{1}(y_i \neq F(x_i))$$

b) Now given this choice of G_m. Optimize over the choice of α. That is solve for

$$\alpha_m = \arg \min_{\alpha} \sum_{i=1}^{n} \exp(-y_i(G(x_i) + \alpha G_m(x_i)))$$

If we define

$$\text{err}_m = \frac{\sum_{i=1}^{n} w_i \mathbb{1}(y_i \neq G_m(x_i))}{\sum_{i=1}^{n} w_i}$$

then you should obtain $\alpha_m = \frac{1}{2} \log\left(\frac{1-\text{err}_m}{\text{err}_m}\right)$

c) The above results yield the following algorithm:

a) Initialize $w_i = \frac{1}{n}$

b) For $m = 1$ to M:

(a) Find

$$G_m = \arg\min_{F \in \mathcal{F}} \sum_{i=1}^{n} w_i \mathbb{1}(y_i \neq F(x_i))$$

(b) Let

$$\text{err}_m = \frac{\sum_{i=1}^{n} w_i \mathbb{1}(y_i \neq G_m(x_i))}{\sum_{i=1}^{n} w_i}$$

(c) Compute $\alpha_m = \frac{1}{2} \log\left(\frac{1-\text{err}_m}{\text{err}_m}\right)$

(d) Set $w_i = w_i \cdot \exp(-\alpha_m y_i G_m(x_i))$

Show that the above procedure is equivalent to the procedure laid out on page 339[1] in the Elements of Statistical Learning Theory book. Note some slight differences. In the above, α_m has a $\frac{1}{2}$ in front. Furthermore, the update for the weights in the ESL has an indicator function, whereas the above does not.

2 Regularization

In later lectures we will discuss the problem of regularization. In this problem we will go over some basic ideas. The aim of regularization is to control the complexity of the underlying model. For decision trees we restrict tree depth. For k-NN we restrict k. In boosting we can control step-size (learning rate), tree depth, and number of total trees to be added.

For standard linear regression problems we can also control the parameters. Two common approaches for regularization are Ridge Regression and Lasso. Ridge-regression aims to control the ℓ_2 norm of the trained parameters. Lasso controls the ℓ_1 norm.

Ridge-regression In Ridge-regression, the optimization problem is to solve

$$\widehat{\theta} = \arg\min_{\theta} \frac{1}{2n} \|X\theta - y\|_2^2 + \lambda \|\theta\|_2^2$$

Note that $\lambda \geq 0$ is a hyper-parameter that we will usually pick with cross-validation. When $\lambda = 0$ we have ordinary least squares. As we increase λ we encourage θ to be closer to 0.

Part a) What is the closed form for the Ridge-regression solution $\widehat{\theta}$?

Part b) Let I be the $p \times p$ identity matrix. What is the closed form solution for Ridge-regression when $X = I$?

[1] https://web.stanford.edu/ hastie/Papers/ESLII.pdf

property of identity matrix: IX=XI=X

X is p × m matrix

Lasso The Lasso is similar. However, instead of an ℓ_2 penalty we have an ℓ_1 penalty. These penalties are known as regularizers that perform regularization.

$$\widehat{\theta} = \arg\min_{\theta} \frac{1}{2n} \|X\theta - y\|_2^2 + \lambda \|\theta\|_1$$

where $\|\theta\|_1 = \sum_{i=1}^{p} |\theta_{(i)}|$. Again, when $\lambda = 0$ we have ordinary least squares and as λ becomes larger the solution is encouraged to be closer to zero. However, due to the change in regularization, the way θ is closer to zero is different. In fact, this penalty is used since it encourages many of the coefficients of the solution to be actually zero. This outcome is desirable for interpretability.

Part c) For general X there is no closed form solution for the Lasso. What is the gradient descent update for this problem? Note that $\|\theta\|_1$ is not differentiable, so you can just take the derivative at 0 to be anything between -1 and $+1$. Take it to be zero for simplicity.

Part d) Suppose that $X = I$. Compute the closed for solution for the Lasso. This time, you will need to use the fact that the derivative of $|s|$ at the point $s = 0$ can be anything between -1 and $+1$. That is the gradient is not unique, but setting one of the possible gradients to be equal to zero is enough to find a solution. Note that you do not need to use gradients to solve this problem, but since the problem is convex you can.