Whiteboard Notes for INDENG 250, Fall 2024

Instructional team: Huiwen Jia, Hao Wang

September 5th 2024

1 Formulas

1.1 Time series

Let $\mathcal{B}$ be the *Lag Operator*, which means given a time series $X = \{X_1, X_2, \dots\}$, $\mathcal{B}^k X_t = X_{t-k}$, where $k \leq t$. For $d = 2$, we have

$$\begin{aligned}(1-\mathcal{B})^2 X_t &= (1 - 2\mathcal{B} + \mathcal{B}^2) X_t, \\ &= X_t - 2X_{t-1} + X_{t-2}.\end{aligned}$$

1.2 Loss/Cost function

Given N samples, the true value and prediction value of sample i is given by y_i and $\hat{y}_i$. The lease square error loss function is defined as:

$$\text{Least Square} = \frac{1}{2} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2.$$

Let $\bar{y}$ represent the sample average. If the restriction is to let $\hat{y}_i, \forall y_i$ to be one single value (for the root, or for a leaf node), then we pick $\hat{y}_i = \bar{y}, \forall y_i$ in this node. And this $\hat{y}_i$ is minimizing the least square error. In this case, the minimized least square error $= 2\times$ deviance, where deviance is defined as

$$\text{Deviance} = \sum_{i=1}^{N} (y_i - \bar{y})^2.$$

1.3 Optimization algorithm

1.3.1 Gradient descent (GD)

Consider the following optimization problem,

$$\begin{aligned} &\min_x f(x) \\ &\text{s.t. } x \in \mathbb{R}^n \end{aligned}$$

Step 0, given initial point x_0, we have function value $f(x_0)$ and corresponding gradient $\nabla f(x_0)$.

Step 1, we want to choose a direction Δx to update x_0, that is, $x_1 = x_0 + \Delta x$. By Taylor expansion, we have

$$f(x_1) \approx f(x_0) + \Delta x \nabla f(x_0)$$

To minimize $f(x_1)$, the optimal direction is given by $\Delta x^* = -\gamma \nabla f(x_0)$, where γ is step size. After we fix the moving direction as $\nabla f(x_0)$, and then the step to find an optimal/good γ is another optimization problem, typically called line search. Instead of finding the optimal γ, another choice is to use a relatively small γ, like 0.1.

1.3.2 GBDT

As we discussed above, after we fit the model in the 0^{th} iteration, the least square loss l is given by

$$l(y, \hat{y}_0) = \frac{1}{2} \sum_{i=1}^{N} (y_i - \hat{y}_{i,0})^2$$

where y_i and $\hat{y}_{i,0}$ are true value and initial prediction of sample i respectively.

In the first iteration, we need to fit a model $g_1(\cdot)$, such that we move our overall prediction output from $\hat{y}_0$ to $\hat{y}_0 + \gamma g_1(\cdot)$. And we hope after the first iteration, we further reduce the loss function, i.e., minimize $l(y, \hat{y}_0 + g_1(\cdot))$. Here we found an analog with gradient descent:

	GD	*GBDT*
objective	$\min\ f(x)$	$\min\ l(y, \hat{y})$
variable	x	$\hat{y}$
current values	x_0	$\hat{y}_0$
outcome of this iteration	$x_1 = x_0 + \Delta x$	$\hat{y}_1 = \hat{y}_0 + \gamma g_1(\cdot)$
what to decide in this iteration	Δx	$\gamma g_1(\cdot)$
given by theory	$\Delta x = -\gamma \nabla f(x_0)$	Similarly, need $g_1(\cdot)$ to fit $-\nabla l(y, \hat{y}_0)$

Then we can derive for i^{th} observation, $\frac{\nabla l}{\hat{y}_{i,\cdot}} = \hat{y}_{i,\cdot} - y_i$. Thus, we evaluate on the current value $\hat{y}_{i,0}$, we have $\frac{\nabla l}{\hat{y}_{i,0}} = \hat{y}_{i,0} - y_i$. By defining pseudo residual(PR) as $y_i - \hat{y}_{i,\cdot}$, the step for fitting the pseudo residual is equivalent to fitting $-\nabla l$.