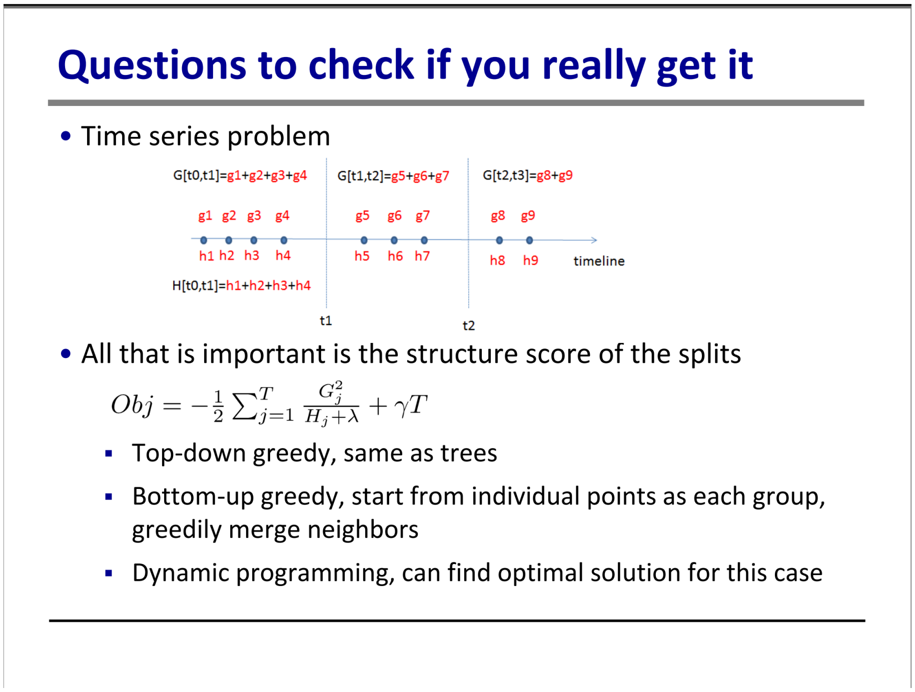
人工智能基础
最佳 split 点的寻找

孔静-2014K8009929022

November 6, 2016

1 问题

“Introduction to Boosted Trees”这个 slides 的 39 页上的问题，提交伪代码，并分析时间复杂度


G[t0,t1]=g1+g2+g3+g4
G[t1,t2]=g5+g6+g7
G[t2,t3]=g8+g9
g1 g2 g3 g4
g5 g6 g7
g8 g9
h1 h2 h3 h4
h5 h6 h7
h8 h9
timeline
H[t0,t1]=h1+h2+h3+h4
t1
t2

Questions to check if you really get it

- Time series problem
- All that is important is the structure score of the splits

$$Obj = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j+\lambda} + \gamma T$$

  - Top-down greedy, same as trees
  - Bottom-up greedy, start from individual points as each group, greedily merge neighbors
  - Dynamic programming, can find optimal solution for this case

ORZ，写完以后想起来要用动态规划写，但好像我用的是分治 >.<。但讲道理差别不大嘛，动态规划用上一次的东西往下传，在这个题里面，分治也可以把这个信息往下传嘛！差别不是很大嘛！懒得改了 =.=

2 伪码

Algorithm 1 Find The Split

```
procedure MAIN(g,h,n,a)
    k = 1
    FindSplit(g,h,n,a)
    return $Split$
end procedure
procedure GETSCORE(G,H,T)
    $Obj = -\frac{1}{2}\sum_{\mathrm{j}} \frac{G[j]^2}{H[\mathrm{j}]+\lambda}+3\gamma$
    return $Obj$                                              ▷ 计算分数
end procedure
procedure GETGAIN(GL,GR,HL,HR)
    $Gain = \frac{1}{2}[\frac{GL^2}{HL+\lambda} + \frac{GR^2}{HR+\lambda} - \frac{(GL+GR)^2}{(HL+HR)+\lambda}] - \gamma$
    return $Gain$                                             ▷ 计算增益
end procedure
procedure FINDSPLIT(g,h,n,a)              ▷ g,h 为 n 个梯度数据，a 为增益阈值
    $G[1] = \sum g[i]$
    $h[1] = \sum h[i]$
    Obj[0] = GetScore(G, H, 1)                     ▷ 用 0 位存储不分割的分数
    Split[0] = 0                              ▷ 用 0 位存储分割点，0 是不分割
    Gain[0] = a                             ▷ 用 0 位暂时存储比较值，a 是阈值
    for i = 1 to n do
        $GR = \sum_{j=1}^{i} g[j]$
        $GR = \sum_{j=i+1}^{n} g[j]$
        $HL = \sum_{j=1}^{i} h[j]$
        $HR = \sum_{j=i+1}^{n} h[j]$
        Gain[i] = GetGain(GL, GR, HL, HR)
        if Gain[i] > Gain[0] then                    ▷ 贪心，找出最大的增益
            Gain[0] = Gain[i]
            Split[0] = i
        end if
    end for
```

Algorithm 2 Find The Split

```
    if Split != 0 then        ▷ 如果找到了分割点，保存数据，分治继续寻找
        Obj[k] = Obj[k-1] - Gain[0]
        Split[k] = Split[0]
        k = k + 1
        for i = 1 to Split do
            gl[i] = g[i]
            hl[i] = h[i]
        end for
        for i = Split + 1 to n do
            gr[i - Split] = g[i]
            hr[i - Split] = h[i]
        end for
        FindSplit(gl, hl, Split, a)
        FindSplit(gr, hr, n - Split, a)
    end if
end procedure
```

3 分析

单层贪心，遍历 O(n)；分治处理多层，O(logn)。
综上，时间复杂度是 O(nlogn)。

4 参考

毕竟英文渣，看了网上翻译版的 =.=
http://www.52cs.org/?p=429
http://blog.sina.com.cn/s/blog_7103b28a0102w6qa.html