به نام خدا

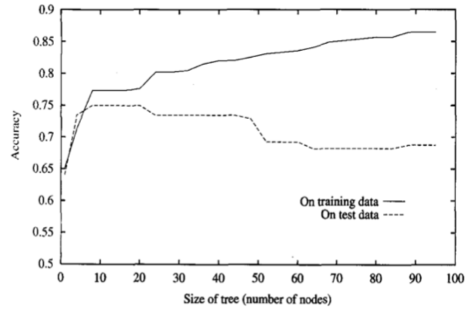
تمرین سوم
درس یادگیری ماشین
دانشکده مهندسی و علوم کامپیوتر
مهلت تحویل : ۲۰ دی ۱۴۰۱ ساعت ۲۳:۵۹
مدرس :دکتر آرمین سلیمی بدر

---

۱. درخت تصمیم:

الف .آیا انتخاب نادرست ریشه‌ی درخت در یک مسئله با برچسب‌های غیر نویزی میتواند باعث شود که درخت تصمیم نتواند به شکل دقیق روی داده‌ها **fit** شود؟ توضیح دهید. (منظور از برچسبهای غیر نویزی آن است که هر نمونه از فضای مسئله، فقط میتواند یک برچسب داشته باشد)

ب .فرض کنید نمودار زیر دقت الگوریتم روی داده‌های آموزش و تست را با بزرگتر شدن درخت تصمیم نمایش میدهد. اگر تعداد داده‌های آموزش را به سمت بی‌نهایت سوق دهیم، این نمودار به چه شکل تغییر خواهد کرد؟



۲. پیاده‌سازی درخت تصمیم:

در این سوال می‌خواهیم دسته‌بند درخت تصمیم را بدون استفاده از کتابخانه‌های سطح بالا پیاده‌سازی کنیم. بدین منظور از یک مجموعه‌ی داده استفاده می‌کنیم که توضیحات و فایل خام داده‌ها در فایل **DT-DATA.ZIP** موجود است. هدف از این تمرین پیاده‌سازی الگوریتم درخت تصمیم از پایه با استفاده از دو معیار **Gini index** و **Information Gain** برای **split** کردن درخت می‌باشد. سپس تاثیر محدود کردن عمق درخت در دقت الگوریتم روی مجموعه داده تست بررسی می‌شود. در نهایت در سلول آخر باید مقایسه‌ای بین حالت عادی و حالتی که عمق درخت محدود شده است ارائه دهید. برای انجام تمرین مراحل مختلف در داخل ژوپیتر همراه آن مشخص شده است. می‌توانید به دلخواه سلولهای دیگری به نوتبوک اضافه کرده و آرگومانهایی به توابع اضافه کنید.

۳. سوال امتیازی (۱۵ نمره):

روش ***Gradient Boosting*** از ترکیب خطی تعدادي مدل ضعیف براي ایجاد یک مدل قوي استفاده میکند. در واقع این روش تلاش میکند با یادگیري ضعیف چندین مدل و ترکیب آنها به یک مدل آموزش دیده قوي برسد. هر مدل ضعیف برخی ضعف‌هاي مدل‌هاي قبلی خود، که هر داده ورودي به شکل سلسله‌مراتبی از آنها عبور کرده تا به مدل ضعیف فعلی رسیده، را میپوشاند و مقدار تابع هزینه را به اندازه توان خودش کاهش میدهد

مراحل مختلف الگوریتم را براي مسئله ***Regression*** و ***Classification*** شرح داده و تفاوت روند محاسبات آنها را بیان کنید.

موفق باشید :)