گروه مهندسی کامپیوتر

رشته تحصیلی: هوش مصنوعی و رباتیکز

نام درس: یادگیری ماشین

تکلیف شماره۲

استاد مربوطه: دکتر مهدی یزدیان

تهیه کننده: حمیدرضا نادمی

بهار ۱۳۹۸

Part A. Theoretical Example

Suppose that instead of selecting a node using information gain (IG) in a binary decision tree, we select a node randomly from nodes with IG>0:

a) Show that each leaf of the tree contains at least one training data.

b) If we have n training data, what is the maximum number of leaf in constructed decision tree? Compare result with the state that we used IG for selecting node.

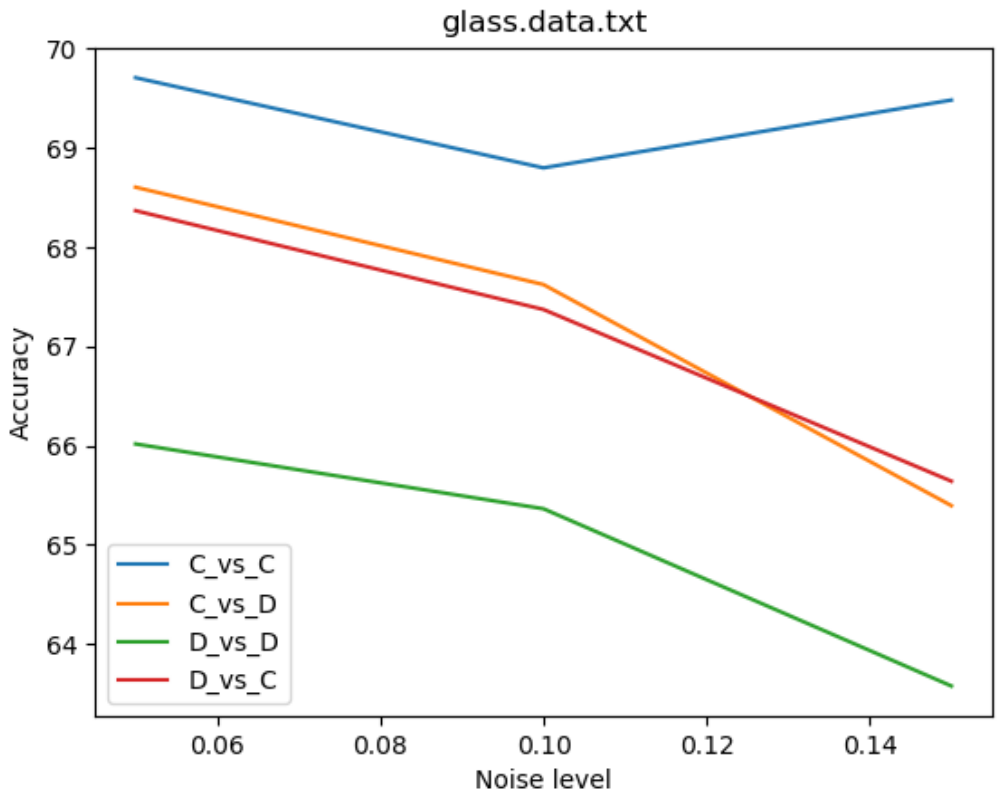
Suppose that you have n non-overlapping points in [0, 1]*[0, 1] space with + and – labels. Also suppose you can select one feature in different levels:

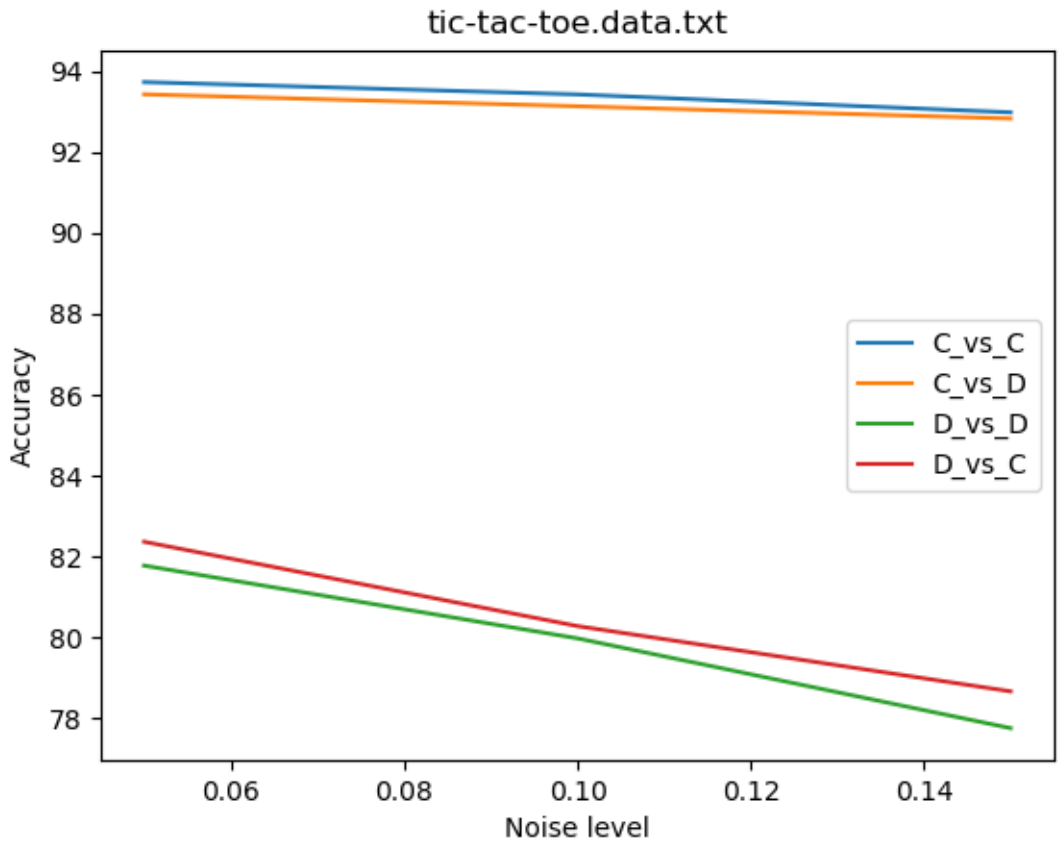
c) Prove that there exist a tree with at least depth of $\log_2 n$ that can classifying data truly.

d) Give an example with n points in defined space that minimal decision tree for them, contains n-1 internal nodes.

Part B. Analysis the effect of attribute noise

1. Plot one figure for each data set that shows the classification accuracy respect to different feature noises with the variance of (5%, 10%, and 15%) of the feature variances. It should be noted that the x-axis and y-axis show noise level and accuracy, respectively. Each figure should contain 4 curves for CvsC, CvsD, DvsC, DvsD results.





2. Analysis the results according to the plots.

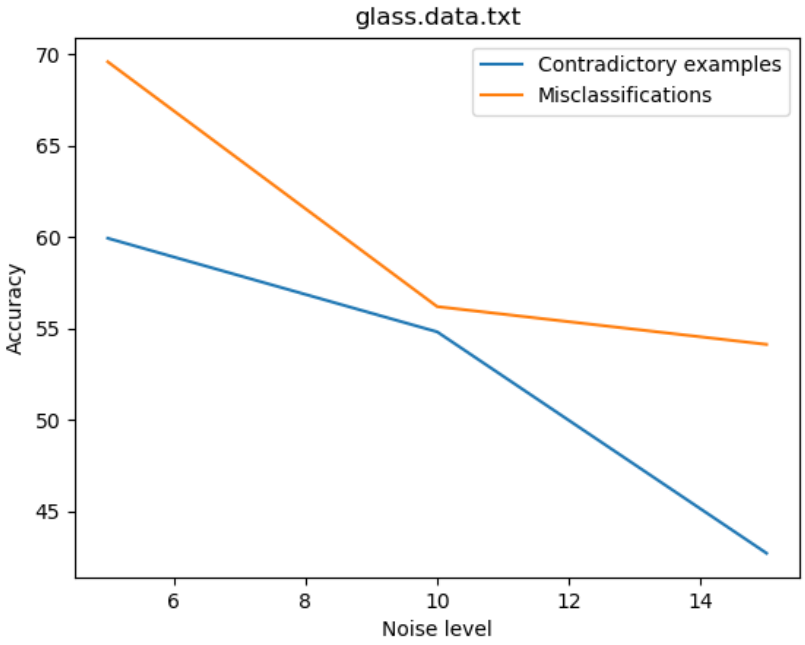
با توجه به شکل۱ و شکل۲، بیشترین دقت کلاس بندی همیشه برای زمانی بوده است که از داده آموزش تمیز و داده تست تمیز (CvsC) استفاده شده است، و هرچه سطح نویز بیشتر شده است دقت دسته بند کمتر شده است (حالات CvsD, DvsC, DvsD) و تفاوتی نداشته است که نویز در داده آموزش بوده یا در داده تست یا هردو.

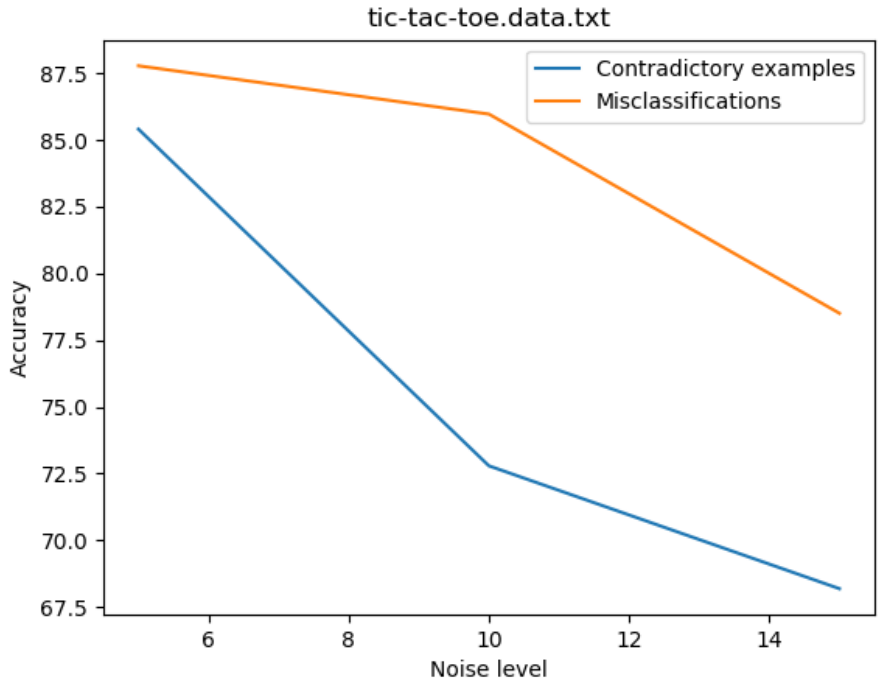
کمترین دقت دسته بند معمولا برای زمانی است که دسته بند با داده‌های آموزش نویزی آموزش دیده و داده‌های تست نویزی را کلاسیفای می‌کند (DvsD). که بیانگر این است که اگر روش های حذف نویز برای دیتاست بکار بگیریم دقت دسته بند بیشتر خواهد شد.

با مقایسه منحنی‌های CvsC و CvsD و مقایسه DvsD و DvsC میبینیم که وجود نویز در داده‌های تست نسبت به زمانی دیتاست تست تمیز است دقت دسته بند کمتر است. (برای دیتاست های واقعی نمیدانیم که داده‌های تست واقعا تمیز هستند یا نویزی و ممکن است نویز در داده‌های تست وجود داشته باشند).

Part C. Analysis the effect of class-label noise

3. Plot one figure for each data set that shows the classification accuracy in terms of different label noise with the level of (5%, 10%, and 15%) of samples. Plot two type of noises over one figure.





4. How do you explain the effect of noise on C4.5 method?

هرچه سطح خطا بیشتر می‌شود دقت دسته بند کمتر می‌شود.

5. In comparison with attribute noise and class noise, which is more harmful? Why?

Class noise اثر بدتری رو دقت دسته بند دارد،