یادگیری ماشین برای بیوانفورماتیک

نیم‌سال دوم ۹۸ـ۹۹

دانشکده‌ی مهندسی کامپیوتر
گردآورند: الهه اکبری، حانیه برغی

تمرین دوم
Ensemble Methods ـ Feature Selection

Ensemble Methods

مسئله‌ی ۱. مقایسه Bagging و Boosting

(۸ نمره) دو دسته‌بند ضعیف داریم. دسته بند اول بر روی داده‌های آموزش overfit شده و برای داده‌های تست عملکرد قابل قبولی ندارد. دسته‌بند دوم بلعکس قدرت تعمیم‌دهی بالایی دارد، اما توانایی یادگیری مدل‌های پیچیده را ندارد. (بایاس در دسته‌بند دوم بالا است.) برای یافتن مدلی با عملکرد مطلوب‌تر می‌توانیم از یکی از دو تکنیک Bagging یا Boosting استفاده کنیم، پیشنهاد شما چیست؟ توضیح دهید.

مسئله‌ی ۲. AdaBoost

آ

(۸ نمره) اگر خطای یادگیرهای ضعیف‌مان در الگوریتم AdaBoost نقض شود و ۵۰ درصد خطا داشته باشیم، چه می‌شود؟ اگر این خطا از ۵۰ درصد بیشتر باشد نیز بیان کنید که الگوریتم چگونه عمل می‌کند و این موضوع در عمل بیانگر چیست؟

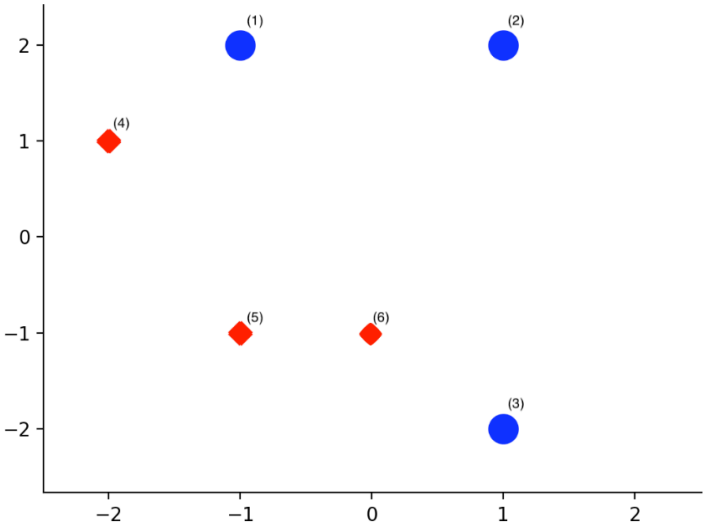
ب

(۱۵ نمره) در یک مسئله دسته‌بندی باینری، فرض کنید n داده آموزش در صفحه داریم که نیمی از نقاط برچسب منفی و نیمی دیگر برچسب مثبت دارند. از نقاطی که برچسب منفی دارند نصف آن‌ها در مختصات $(۱, ۱)$ و بقیه نقاط منفی در مختصات $(-۱, -۱)$ قرار گرفته اند. از نقاط مثبت، $\frac{m(۱-\epsilon)}{۴}$ نقطه در مختصات $(-۱, ۱)$ و $\frac{m(۱+\epsilon)}{۴}$ نقطه دیگر در مختصات $(۱, -۱)$ قرار گرفته اند. ϵ یک عدد ثابت بین ۰ و $\frac{۱}{۲}$ است. رفتار الگوریتم AdaBoost روی این نمونه‌ها را شرح دهید و بیان کنید که جواب این الگوریتم بعد از T گام چیست؟

پ

(۹ نمره) می‌خواهیم داده‌های زیر را دسته‌بندی کنیم. فضای فرضیه برای دسته‌بندهای ضعیف را یک خط افقی یا عمودی در نظر بگیرید. الگوریتم AdaBoost را تا ۲ مرحله اجرا کنید:

- در هر مرحله مقادیر مربوط به α_t، ϵ_t، Z_t و $d_t(i)$ را بدست آورید.
- در نهایت خطای مربوط به اجرای الگوریتم را بدست آورید.



ت

(۱۵ نمره) در یک مسئله دسته‌بندی با استفاده از روش AdaBoost می‌خواهیم دسته‌بندی برای داده‌های آموزش $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ پیدا کنیم. اگر تابع هزینه این مسئله برابر ($[TC]$ مجموعه داده‌های درست دسته‌بندی شده و $[MC]$ مجموعه داده‌های اشتباه دسته‌بندی شده است.)

$$E = e^{-\alpha_t/2} \sum_{i \in [TC]} \omega_i^{(t)} + e^{\alpha_t/2} \sum_{i \in [MC]} \omega_i^{(t)}$$

باشد.
۱. ثابت کنید وزن دسته‌بند در مرحله tام با استفاده از رابطه $\alpha_t = ln(\frac{1-\epsilon_t}{\epsilon_t})$ به‌روزسانی می‌شود. که در این رابطه ϵ_t برابر است با:

$$\epsilon_t = \frac{\sum_{i=1}^{n} \omega_i^{(t)} I(f_t(x_i) = y_i)}{\sum_{i=1}^{n} \omega_i^{(t)}}$$

توجه کنید که $\omega_i^{(t)}$ وزن داده iام در گام tام را نشان می‌دهد. و $f_t(.)$ دسته‌بند مرحله tام است.
۲. تحقیق کنید این تابع هزینه چرا خوب عمل می‌کند و چه استفاده‌ای از آن می‌توانیم بکنیم؟

مسئله‌ی ۳. Random Forest

آ

(۶ نمره) برای یک مسئله دسته‌بندی از Random Forest استفاده کردیم، اما جواب خوبی بدست نیامد. برای بهبود این روش دو پیشنهاد داریم: ۱. افزایش عمق درخت‌ها. ۲. افزایش تعداد درخت‌ها. این دو روش برای بهبود را با یکدیگر مقایسه کنید و بیان کنید هر کدام در چه شرایطی می‌توانند مفید باشند؟

ب

(۶ نمره) توضیح دهید چرا Random Forest این امکان را به ما می‌دهد تا با ثابت ماندن عمق درخت‌ها، عملکرد بهتری در دسته‌بندی داشته باشیم.

Feature Selection

مسئله‌ی ۴. Wrapper

(۶ نمره) چرا روش‌های Wrapper در برابر بیش‌برازش مقاوم هستند؟

مسئله‌ی ۵. Filter

(۵ نمره) در مورد معیارهای فیلتر تک بعدی، درستی جمله زیر را بررسی کنید:
«بسیاری از معیارهای فیلتر تک بعدی عملا معادل با کیفیت دسته‌بندی یا (رگرسیون) داده‌ها با استفاده از آن ویژگی هستند.»

مسئله‌ی ۶. Markov Blanket

(۱۵ نمره) در این سوال، هدف اثبات یکی از ویژگی‌های Markov Blanket در شبکه‌های بیزین است. در یک شبکه بیزین، با فرض داشتن Markov Blanket ثابت کنید: «هر گره در شبکه بیزین از بقیه زیرمجموعه از گره‌ها مستقل است.»
یادآوری: Markov Blanket یک گره در شبکه‌های بیزین از سه گره ،parents children و coparents تشکیل شده است.

مسئله‌ی ۷. Lasso

(۷ نمره) برای یافتن ویژگی‌های مهم در یک مسئله رگرسیون ابعاد بالا روش‌های متفاوتی وجود دارد، اما یکی از مهم‌ترین و پرکاربردترین این روش‌ها استفاده از تابع هزینه Lasso است. این تابع هزینه را بنویسید و بررسی کنید این روش چه ویژگی مهمی در استخراج ویژگی‌های مهم دارد که پرطرف‌دار شده است؟ سپس این تابع هزینه را با روش Ridge Regression[1] مقایسه کنید. آیا این روش برای استخراج ویژگی‌ها مناسب است؟

[1] یک روش برای رگرسیون با تابع هزینه مجموع مربعات خطا و جمله منظم‌ساز نرم ۲ است.