دانشکده‌ی مهندسی کامپیوتر

یادگیری ماشین بهار ۱۴۰۰

تمرین سری هفتم

مدرس: دکتر محمّدحسین رهبان زمان تحویل: ۱۹ خرداد

سوال ۱ *AdaBoost Algorithm*

آ) همانطور که می‌دانیم در هر مرحله از الگوریتم *Adaboost*، یک دسته‌بند که کمترین خطا را با در نظر گرفتن توزیع آن مرحله دارد، انتخاب می‌شود. اثبات کنید این الگوریتم هیچگاه دو تابع یکسان را در دو مرحله متوالی انتخاب نمی‌کند ($h_t \neq h_{t+1}$). (۱۰ نمره)

ب) بردار توزیع $\big(\mathcal{D}_{t+1}(1), \mathcal{D}_{t+1}(2), \ldots, \mathcal{D}_{t+1}(m)\big)$ را در الگوریتم *Adaboost* در نظر بگیرید که m بیانگر تعداد داده‌هاست. اثبات کنید این بردار و برداری که مولفه‌های آن $y_i h_t(\mathbf{x}_i)$ها هستند، ناهمبسته[1]اند (به این معنی که ضرب داخلی آن‌ها صفر است). (۱۰ نمره)

سوال ۲ *Density Estimation with Non–Parametric Methods*

مجموعه داده $X_1, X_2, \ldots, X_n$ را در نظر بگیرید. می‌خواهیم تابع چگالی احتمال این داده‌ها را تخمین بزنیم.

آ) یک روش پارامتری برای این مساله ارایه داده و مزایا و معایب آن را بیان کنید. (۵ نمره)

در ادامه، این مساله را از جنبه‌ی دیگری بررسی می‌کنیم. همانطور که پیش‌تر نیز گفتیم، هدف ما تخمین تابع چگالی احتمال این داده‌هاست. برای سادگی فرض کنید دامنه‌ی این توزیع، بازه‌ی $[0,1]$ است و داریم $\forall \mathbf{x} \in [0,1] : |p'(\mathbf{x})| \leq \beta$ که p بیانگر تابع چگالی احتمال است. برای حل مسئله به روش بافتنگار[2]، فرض کنید بازه‌ی $[0,1]$ را به H بازه‌ی مساوی $I_1, I_2, \ldots, I_H$ تقسیم کنیم:

$$I_1 = \big[0, \frac{1}{H}\big),\ I_2 = \big[\frac{1}{H}, \frac{2}{H}\big),\ \ldots,\ I_H = \big[\frac{H-1}{H}, 1\big]$$

حال برای محاسبه چگالی داده‌ی جدید $\mathbf{x} \in I_l$ داریم:

$$\hat{p}(\mathbf{x}) = \frac{\textbf{number of observations within } I_l}{n} \times \frac{1}{\frac{1}{H}} = \frac{H}{n} \sum_{i=1}^{n} \mathbb{1}_{(\mathbf{x}_i \in I_l)}$$

ب) نابراری زیر را اثبات کنید: (۱۵ نمره)

$$\textbf{bias}\big(\hat{p}(\mathbf{x})\big) \leq \frac{\beta}{H}$$

[1] Uncorrelated

[2] Histogram

پ) از رابطه‌ی بدست آمده در بخش **ب** چه نتیجه‌ای می‌گیرید؟ استدلال کنید که چرا این روش *Non–Parametric* است؟ (۵ نمره)

ت) روش بخش **ب** را با روش پیشنهادی بخش **آ** مقایسه کنید. (۵ نمره)

سوال ۳ *Ensemble Learning*

یک مسئله‌ی دسته‌بندی دودویی را در نظر بگیرید. فرض کنید T دسته‌بند داریم که هریک به صورت مستقل، خطای تعمیم‌پذیری[۳] ϵ دارند. به عبارتی داریم $\mathbb{P}\big(h_i(\mathbf{x}) \neq f(\mathbf{x})\big) = \epsilon$ که f نشان دهنده *Ground Truth Function* است. تابع H را به شکل زیر تعریف می‌کنیم:

$$H(\mathbf{x}) = \boldsymbol{sign}\Big(\sum_{i=1}^{T} h_i(\mathbf{x})\Big)$$

آ) توضیح دهید H در چه صورتی مرتکب خطا می‌شود؟ (۵ نمره)

ب) یک حد بالا برای احتمال رخدادن خطا بدست آورده و نشان دهید با افزایش T به سمت بی‌نهایت، احتمال رخدادن خطا به سمت صفر می‌رود. (۱۰ نمره)

راهنمایی: می‌توانید از نامساوی هافدینگ استفاده کنید.

سوال ۴ (عملی) فایل *Notebook*ی که در اختیار شما قرار داده شده را کامل کنید. در این سوال می‌خواهیم ۳ مدل رگرسیون خطی[۴]، رگرسیون ناپارامتری[۵] و نزدیک‌ترین همسایه[۶] را روی مجموعه دادگانی که در اختیار شما قرار می‌گیرد، پیاده سازی کرده و نتیجه‌ها را بررسی کنید. (۳۰ نمره)

پاینده باشید

[3] Generalization Error
[4] Linear Regression
[5] Nonparametric Regression
[6] Nearest Neighbors