یادگیری ماشین

نیم‌سال اول ۱۴۰۰-۱۴۰۱

مدرس: دکتر عباس حسینی

دانشکده مهندسی کامپیوتر

تمرین سری چهارم

دسته‌بندها

موعد تحویل: ۵ آذر

پرسش ۱

از معروف‌ترین و مهم‌ترین دسته‌بندهای خطی، دسته‌بند Perceptron می‌باشد که شبه‌کد آن در زیر آمده‌است.

الگوریتم Perceptron یک الگوریتم تکرار شونده است که توالی از بردارهای $w_1, w_2, \ldots$ را می‌سازد. به عنوان مقداردهی اولیه، w_1 را بردار تمام صفر قرار می‌دهیم. در مرحله‌ی t، اگر داده‌ی i وجود داشته باشد که به طوری که توسط w_t اشتباه دسته‌بندی شود یا به عبارت دقیقتر $\text{sign}(\langle w_t, x_i \rangle) \neq y_i$ باشد، آنگاه مرحله‌ی بروزرسانی بردار w_t به صورت $w_{t+1} \leftarrow w_t + y_i x_i$ انجام می‌شود. قاعدتا هدف نهایی ما این است که به w_T برسیم به گونه‌ای که $\forall i \; : \; y_i \langle w_T, x_i \rangle > 0$ باشد، یعنی هیچ اشتباهی روی داده‌هایی که در اختیار داریم نداشته باشد. برای سادگی فرض می‌کنیم داده‌هایی که در اختیار داریم با دسته‌بند خطی، قابل جدا شدن هستند. حالت کلی و درنظر نگرفتن این شرط، مساله‌ی جالب دیگری است که پیشنهاد می‌کنم روی آن فکرکنید!

Batch Perceptron

input: A training set $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$
initialize: $\mathbf{w}^{(1)} = (0, \ldots, 0)$
  for $t = 1, 2, \ldots$
    if $(\exists\, i \;\text{ s.t. }\; y_i \langle \mathbf{w}^{(t)}, \mathbf{x}_i \rangle \leq 0)$ **then**
      $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + y_i \mathbf{x}_i$
    else
      output $\mathbf{w}^{(t)}$

الف) پارامتر B را به صورت $B = \min\{\|w\|_2 : \forall i \in [m], y_i \langle w, x_i \rangle \geq 1\}$ و همچین پارامتر R را به صورت $R = \max_i \|x_i\|_2$ تعریف می‌کنیم. اثبات کنید که این الگوریتم حداکثر $(RB)^2$ مرتبه اجرا می‌شود.

ب) اگر مرحله‌ی بروزرسانی را به صورت $w_{t+1} \leftarrow w_t + \eta y_i x_i$ انجام دهیم که η یک ابرپارامتر است که عددی نامنفی است، اثبات کنید تعداد مراحلی که تا همگرایی طی می‌شود برای هر دو حالت الگوریتم یکسان خواهد بود.

پرسش ۲

الگوریتم Adaboost را در نظر بگیرید. شبه‌کد این الگوریتم در زیر آمده است.

$\textsc{AdaBoost}(S = ((x_1, y_1), \ldots, (x_m, y_m)))$

for $i \leftarrow 1$ **to** m **do**
  $\mathcal{D}_1(i) \leftarrow \frac{1}{m}$
for $t \leftarrow 1$ **to** T **do**
  $h_t \leftarrow$ base classifier in $\mathcal{H}$ with small error $\epsilon_t = \mathbb{P}_{i \sim \mathcal{D}_t}\left[h_t(x_i) \neq y_i\right]$
  $\alpha_t \leftarrow \frac{1}{2} \log \frac{1-\epsilon_t}{\epsilon_t}$
  $Z_t \leftarrow 2\left[\epsilon_t(1-\epsilon_t)\right]^{\frac{1}{2}}$ ▷ normalization factor
  for $i \leftarrow 1$ **to** m **do**
    $\mathcal{D}_{t+1}(i) \leftarrow \frac{\mathcal{D}_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t}$
$f \leftarrow \sum_{t=1}^{T} \alpha_t h_t$
return f

الف) اگر دقت کنید در الگوریتم Adaboost در هر مرحله یک دسته‌بند به صورتی که طبق توزیع آن مرحله کمترین خطا را داشته باشد انتخاب می‌شود. اثبات کنید این الگوریتم هیچگاه دو تابع یکسان در دو مرحله متوالی انتخاب نمی‌کند ($h_t \neq h_{t+1}$).

ب) بردار توزیع $(D_{t+1}(1), D_{t+1}(2), ..., D_{t+1}(m))$ را در الگوریتم Adaboost درنظر بگیرید که m تعداد داده‌ها است. اثبات کنید این بردار و برداری که مولفه‌های آن $y_i h_t(x_i)$ هستند، uncorrelated هستند (به این معنی که ضرب داخلی آن‌ها صفر است).

پرسش ۳

یک مساله‌ی دسته‌بندی باینری را در نظر بگیرید. هر دسته‌بند h در چارجوب این مساله، تابعی از $\mathcal{X}$ به $\{0,1\}$ یعنی $h : \mathcal{X} \to \{0,1\}$ است. نرخ خطا را به صورت زیر تعریف می‌کنیم.

$$\mathrm{R(h)} = \mathbf{P}[Y \neq \mathrm{h}(X)]$$

و قاعدتا به همین ترتیب نرخ خطای تجربی نیز به صورت زیر تعریف می‌شود.

$$\hat{\mathrm{R}}(\mathrm{h}) = \frac{1}{n}\sum_{i} \mathbf{I}(\mathrm{h}(X_i) \neq Y_i)$$

الف) اثبات کنید که تابع hای که $\mathrm{R(h)}$ کمینه کند به صورت زیر است.

$$\mathrm{h}^*(x) = \begin{cases} 1 & \text{if } \mathrm{m}(x) > \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

که تابع m برابر $\mathrm{m}(x) = \mathbb{E}(Y \mid X = x) = \mathbb{P}(Y = 1 \mid X = x)$ است.
حال فرض کنید که $\hat{\mathrm{m}}$ تخمینی از m باشد و مطابق با آن $\hat{\mathrm{h}}$ را به صورت زیر تعریف کنیم.

$$\widehat{\mathrm{h}}(x) = \begin{cases} 1 & \text{if } \widehat{\mathrm{m}}(x) > \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

ب) با فرض اینکه $R^* = R(h^*)$ است، اثبات کنید

$$R(\widehat{h}) - R^* \leq 2\int |\widehat{m}(x) - m^*(x)|\, P_X(x)dx$$

پرسش ۴

می‌خواهیم با استفاده از درخت تصمیم‌گیری، دسته‌بندی طراحی کنیم که برحسب ویژگی‌های هر رستوران پیشبینی کند که آیا از رستوران رضایت خواهیم داشت یا نه. داده‌هایی که جمع‌آوری کرده‌ایم را در شکل زیر مشاهده می کنید.

رضایت‌مندی	محدودیت‌ها	محل رستوران	رنج قیمت	نوع غذا	رستوران
0	Vegeterian	غرب شهر	ارزان	فست فود	1
0	Gluten free	شمال شهر	متوسط	ایرانی	2
0	هیچی	شمال شهر	متوسط	آسیایی	3
0	Vegeterian	شرق شهر	گران	آسیایی	4
1	Vegeterian	غرب شهر	ارزان	آسیایی	5
1	هیچی	شمال شهر	متوسط	فست فود	6
1	هیچی	شمال شهر	ارزان	ایرانی	7
0	Gluten free	شرق شهر	ارزان	آسیایی	8
0	هیچی	غرب شهر	گران	فست فود	9
1	Vegeterian	شرق شهر	متوسط	ایرانی	10
1	Gluten free	شرق شهر	متوسط	آسیایی	11

الف) با استفاده از داده‌های داده‌شده، درخت تصمیم‌گیری رضایت نداشتن یا نداشتن از رستوران را بسازید. سعی کنید کل مراحل روند حل را توضیح دهید.

ب) حال با استفاده از دسته‌بندی که ساخته‌اید، رضایت‌مندی پنج رستوران زیر را پیش‌بینی کنید.

محدودیت‌ها	محل رستوران	رنج قیمت	نوع غذا	رستوران
هیچی	شمال شهر	ارزان	فست فود	12
هیچی	شرق شهر	متوسط	ایرانی	13
Gluten free	غرب شهر	ارزان	ایرانی	14
Vegeterian	شرق شهر	ارزان	آسیایی	15
Gluten free	شمال شهر	ارزان	ایرانی	16

پ) اگر بدانیم مقدار واقعی رضایت به ترتیب به صورت $[0, 1, 0, 1, 0]$ باشد، F_1 score را محاسبه کنید.

پرسش ۵

یک مدل دسته‌بند مولد برای K کلاس را در نظر بگیرید به طوری که توزیع پیشین روی کلاس‌ها به صورت $p(\mathcal{C}_k) = \pi_k$ باشد و احتمال متعلق بودن به دسته‌ی k برحسب ویژگی‌ها به صورت $P(\boldsymbol{\phi} \mid \mathcal{C}_k)$ نشان دهیم که ϕ ویژگی‌های داده است. مجموعه‌ی داده‌هایی که در اختیار داریم به صورت زوج‌های $\{\boldsymbol{\phi}_n, \mathbf{t}_n\}$ هستند که t_n بردار باینری K بعدی است که اگر متعلق به دسته‌ی j باشد، مولفه‌ی j آن 1 و باقی مولفه‌های آن 0 است. همچنین فرض کنید داده‌ها به صورت مستقل تولید شده‌اند.

الف) نشان دهید جواب بیشینه درست نمایی برای توزیع پیشین به صورت زیر خواهد بود.

$$\pi_k = \frac{N_k}{N}$$

که N تعداد کل داده‌ها و N_k تعداد داده‌های متعلق به دسته‌ی k است.

حال فرض کنید توزیع ویژگی داده نسبت به کلاس به صورت زیر باشد

$$P(\boldsymbol{\phi} \mid \mathcal{C}_k) = \mathcal{N}(\boldsymbol{\phi} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma})$$

ب) تخمین بیشینه درست نمایی برای پارامترهای μ_k و Σ را به دست آورید.

پرسش ۶

سوال عملی

در این سوال می‌خواهیم الگوریتم Adaboost را که شبه‌کد آن در صفحه‌ی قبل نیز آمده، روی یک مجموعه‌ی داده‌ای که به صورت خطی جداپذیر نیست بررسی کنیم. این مجموعه داده، در فایل data.csv قرار گرفته است.

الف) اگر دقت کنید برای هر داده، ۲ ویژگی در نظر گرفته شده و هر داده متعلق به یکی از دو کلاس است. ابتدا نمودار این داده‌ها را رسم کنید به گونه‌ای که داده‌های متعلق به یک کلاس رنگ یکسانی داشته باشند. همچنین تقسیم بندی داده‌های تست و آموزش را نیز در این قسمت انجام دهید.

ب) نکته‌ی مهم دیگر این الگوریتم، انتخاب مجموعه‌ی توابع فرض یعنی H است. برای راحتی می‌توانید از کتابخانه‌ی sklearn برای این بخش استفاده کنید. به طور مثال از کلاس DecisionTreeClassifier آن و یا هر کلاس دیگری جز کلاس مربوط به Boosting انتخاب کنید. پس از انتخاب H مراحل یادگیری را مطابق الگوریتم طی کرده تا به تابع نهایی برسید.

پ) دقت (accuracy) تابع به دست آمده را روی مجموعه داده‌ی تست محاسبه کنید.

داده‌ها از اینجا قابل دریافت است.