# بسمه تعالی

دانشکده مهندسی کامپیوتر

درس یادگیری ماشین
نیم‌سال دوم ۰۴-۰۳
استاد: ربیعی_ دهقانیان

امتحان پایان‌ترم | زمان امتحان: ۳ ساعت

## بخش مفهومی (۳۰ نمره)

برای هر یک از این سوالات این بخش، در یک پاراگراف توضیحات خواسته شده را ارائه دهید. نمره‌ی تمام سوالات این بخش با یکدیگر برابر است.

۱.۱ فرض کنید در یک مجموعه آموزشی، تمامی ویژگی‌های داده‌ها **عددی و پیوسته** هستند. دو دانشجو الگوریتم درخت تصمیم را با دو رویکرد متفاوت اجرا می‌کنند:

- دانشجوی اول، الگوریتم درخت تصمیم را **مستقیماً روی داده‌های خام و بدون هیچ پیش‌پردازشی** اجرا می‌کند و درخت $T_1$ را به‌دست می‌آورد.
- دانشجوی دوم، پیش از اجرای الگوریتم، داده‌های عددی را **استانداردسازی می‌کند** (با کم‌کردن میانگین و تقسیم بر انحراف معیار هر ویژگی) و سپس الگوریتم را اجرا کرده و درخت $T_2$ را به‌دست می‌آورد.

با توجه به تأثیر پیش‌پردازش بر عملکرد مدل‌های درختی، دو درخت حاصل را از نظر **ساختار (شکل)** با یکدیگر مقایسه کنید.

۲.۱ یکی از مشکلات درخت تصمیم این است که در عین حال که به دقت بالایی نسبت به داده می‌رسد اما واریانس خطای آن بالاست. توضیح دهید که جنگل تصادفی چگونه با وجود حفظ دقت بالا، واریانس خطای آن را کاهش می‌دهد.

۳.۱ در روش PCA که در کلاس صحبت شده است بدین صورت عمل می‌کنیم که ابتدا داده‌های آموزش را نرمال می‌کنیم و سپس بردارهای اصلی را می‌یابیم و داده‌های آموزش را در این فضای جدید بررسی می‌کنیم. برای داده‌های تست نیز نرمال‌سازی داده‌های آموزش را اعمال کرده و سپس داده‌ها را به فضای برداری جدید تصویر می‌کنیم. دانشجویی پیشنهاد داده است از آن‌جایی که PCA رویکردی بدون نظارت[1] می‌باشد، می‌توانیم قبل از این‌که داده‌ها را به دو بخش آموزش و تست تقسیم کنیم، نرمال‌سازی و یافتن بردارهای اصلی را بر روی تمام داده ها (آموزش و تست) انجام دهیم و پس از انجام این کار، داده‌ها در فضای جدید را به آموزش و تست تقسیم کنیم و فرآیند آموزش را انجام دهیم. صحت پیشنهاد او را بررسی کنید.

۴.۱ بررسی کنید که اگر به همه نمونه های ماتریس ویژگی‌های $X$ یک بعد با مقدار ثابت ۱ اضافه کنیم تغییري در نتیجه PCA ایجاد می شود یا خیر.

۵.۱ بررسی کنید اگر از روش PCA براي کاهش بعد مسئله از $D$ به $K$ استفاده کنیم، آیا خروجی شبیه حالتی است که ابتدا این روش را برای کاهش بعد مسئله از $D$ به $L(L > K)$ به‌کار ببریم و سپس ویژگی های ماتریس جدید را با استفاده از این روش از $L$ به $K$ کاهش بعد دهیم یا خیر.

۶.۱ اگر از روش bootstraping استفاده کنیم تا $N' = Np$ داده از $N$ داده‌ی موجود نمونه‌گیری کنیم، ثابت کنید به‌طور تقریبی $e^{-p}.N$ از داده‌ها به‌طور کلی در نمونه‌گیری انتخاب نخواهند شد. (فرض کنید که $N$ بسیار بزرگ است)

[1] unsupervised

## بخش تئوری (۷۰+ ۱۵ نمره)

۲.(۲۵ نمره) فرض کنید $n \geq ۳$ داده در فضای دو بعدی داریم که هر کدام به یکی از دو کلاس $+۱$ یا $-۱$ تعلق دارند (تضمین می‌شود که از هر کلاس حداقل یک داده داریم). هم‌چنین می‌دانیم که این داده‌ها به صورت خطی تفکیک‌پذیر هستند. قصد داریم با استفاده از رویکرد ماشین بردار پشتیبان سخت (hard SVM) ابرصفحه‌ی جداکننده‌ی این نقاط که در این‌جا یک خط می‌باشد را پیدا کنیم. همان‌طور که می‌دانید این مسئله را در حالت کلی می‌توان به صورت مسئله‌ی بهینه‌سازی مقید زیر بیان کرد:

$$\min_{\mathbf{w}} \frac{۱}{۲}\|\mathbf{w}\|^۲ \quad \text{subject to} \quad y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq ۱, \quad i = ۱, ۲, \ldots, N.$$

اثبات کنید که این مسئله حداقل ۳ بردار پشتیبان دارد.

**توجه:** در صورتی که بتوانید ثابت کنید این مسئله حداقل ۲ بردار پشتیبان دارد، ۱۵ نمره از ۲۵ نمره‌ی سوال به شما تعلق خواهد گرفت.

۳.(۲۰ نمره) فرض کنید $X = x_۱, x_۲, \ldots, x_n$ داده‌های ما باشد و $\gamma$ یک ماتریس Indicator باشد به این صورت که $\gamma_{ij} = ۱$ اگر $x_i$ متعلق به خوشه $j$ام باشد و در غیر این صورت برابر ۰ است. فرض کنید $\mu_۱, \ldots, \mu_k$ میانگین خوشه‌ها باشند.
اعوجاج $J$ برای داده‌ها به صورت زیر محاسبه می‌شود:

$$J(\gamma, \mu_۱, \ldots, \mu_k) = n\sum_{j=۱}^{k}\sum_{i=۱}^{n}\gamma_{ij}\|x_i - \mu_j\|^۲$$

همچنین $C = ۱, \ldots, k$ را به عنوان مجموعه خوشه‌ها در نظر بگیرید.

با توجه به الگوریتم Kmeans به سوالات زیر جواب دهید.

(الف) نشان دهید که الگوریتم در تعداد متناهی قدم به پایان می‌رسد.

(ب) فرض کنید $\hat{x}$ میانگین داده‌های نمونه باشد. مقادیر زیر را در نظر بگیرید.

$$T(X) = \frac{\sum_{i=۱}^{n}\|x_i - \hat{x}\|^۲}{n}$$

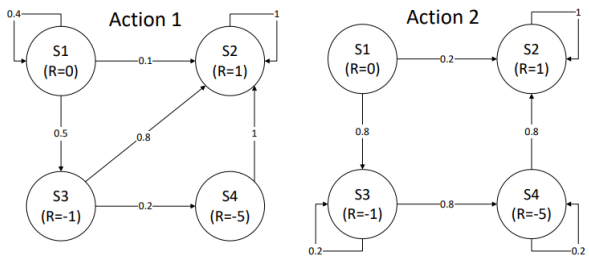
$$W_j(X) = \frac{\sum_{i=۱}^{n}\gamma_{ij}\|x_i - \mu_j\|^۲}{\sum_{i=۱}^{n}\gamma_{ij}}$$

$$B(X) = \sum_{j=۱}^{k}\frac{\sum_{i=۱}^{n}\gamma_{ij}}{n}\|\mu_j - \hat{x}\|^۲$$

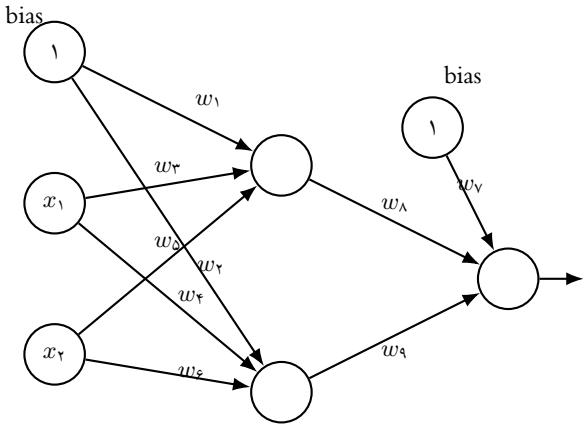
در اینجا، $T(X)$ نشان‌دهنده‌ی انحراف کلی، $W_j(X)$ انحراف درون خوشه‌ای و $B(X)$ انحراف بین خوشه‌ای است. رابطه بین این ۳ مقدار به چه صورت است؟ نشان دهید که K-means می‌تواند به عنوان کمینه کننده‌ی میانگین وزن دار مقادیر درون خوشه‌ای و به طور تقریبی بیشینه کردن انحراف بین خوشه‌ای دیده شود.

۴.(۲۰ نمره) با توجه به فرآیند تصمیم‌گیری مارکوف (Markov Decision Process) در شکل، مقادیر مربوط به هر حالت (تابع ارزش یا Value Function) را به دست آورید. توجه کنید که حالت‌ها به ازای هر حرکت دوبار تکرار شده‌اند تا از پیچیدگی شکل کاسته شود. روی هر یال، احتمال انتقال از حالت مبدأ به مقصد در صورت انتخاب اکشن مربوطه نوشته شده است.

همچنین پاداش هر حالت (که بین دو اکشن مقدار ثابتی دارد و در هنگام خروج از حالت دریافت می‌شود) با نماد $R$ روی هر حالت نشان داده شده است. از $\gamma = ۰/۹$ استفاده کنید.




۵. (۲۰ نمره) شبکه عصبی زیر را در نظر بگیرید که برای حل یک مسئله دسته‌بندی دوکلاسه طراحی شده و دارای یک لایه مخفی است. در نورون‌های لایه مخفی، از تابع فعال‌ساز خطی $h(z) = cz$ استفاده شده و برای نورون خروجی، تابع فعال‌ساز سیگموید $g(z) = \frac{۱}{۱+e^{-z}}$ به کار رفته است تا تابع $P(y = ۱|x, \mathbf{w})$ یاد گرفته شود که در آن $x = (x_۱, x_۲)$ و $\mathbf{w} = (w_۱, w_۲, \ldots, w_۹)$ می‌باشد.




(الف) خروجی شبکه را بر حسب ورودی، وزن‌ها و بایاس‌ها بنویسید. سپس مرز تصمیم این شبکه عصبی را مشخص کنید.
(ب) یک شبکه عصبی بدون لایه مخفی رسم کنید که خروجی آن معادل شبکه شکل فوق باشد. وزن‌های این شبکه را بر حسب وزن‌های شکل فوق مشخص کنید.
(ج) آیا ممکن است تعداد نورون‌ها یا حذف لایه مخفی در شبکه اولیه، باعث تغییر در بایاس و واریانس شده است؟
(د) آیا می‌توان گفت که هر شبکه عصبی چندلایه که با تابع فعال‌ساز خطی قابل تبدیل به یک شبکه عصبی بدون لایه مخفی است؟ مختصر توضیح دهید.

- سوال امتیازی از مباحث میان‌ترم

در مسئله رگرسیون خطی، قصد داریم به نمونه‌های آموزشی، وزن‌های متفاوتی نسبت دهیم. به بیان دقیق‌تر، می‌خواهیم مقدار $J(\theta)$ را کمینه کنیم که به صورت زیر تعریف می‌گردد:

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m} w(i)\left(\theta^T x^{(i)} - y^{(i)}\right)^2$$

(الف) نشان دهید ماتریس $W$ موجود است، به طوری که داریم:

$$J(\theta) = (X\theta - y)^T W (X\theta - y)$$

(ب) با محاسبه $\nabla_\theta J(\theta)$ و برابر قرار دادن آن با صفر، مقدار $\theta$ ای را که $J(\theta)$ را کمینه می‌کند، بیابید. (توجه: در حالتی که همه وزن‌ها یکسان باشند، می‌دانیم $\theta^* = (X^T X)^{-1} X^T y$. جواب‌تان برای این قسمت باید یک فرم بسته باشد که تابعی از $W$، $X$ و $y$ است.)

(ج) فرض کنید مجموعه داده $\{(x^{(i)}, y^{(i)}) : i = 1, 2, \ldots, m\}$ شامل نمونه مستقل داده شده است. قصد داریم $y^{(i)}$ها را گونه‌ای مدل کنیم که گویی از توزیع‌های شرطی با سطوح مختلفی از واریانس گرفته شده‌اند. به طور مشخص، فرض کنید داریم:

$$p(y^{(i)}|x^{(i)};\theta) = \frac{1}{\sqrt{2\pi\sigma(i)}} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma(i)}\right)$$

به بیان دیگر، از یک توزیع گاوسی با میانگین $\theta^T x^{(i)}$ و واریانس $\sigma(i)$ می‌آید. $\sigma(i)$ها ثابت هستند و مقدارشان مشخص است. نشان دهید به لحاظ ریاضی، تخمین بیشینه درست‌نمایی برای $\theta$، معادل است با حل یک مسئله رگرسیون خطی وزن‌دار. به طور مشخص مقادیر $\sigma(i)$ها را بر حسب $w(i)$ به‌دست آورید.