دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

# یادگیری ماشین

**پاییز ۱۴۰۳**

استاد: علی شریفی زارچی
مسئول تمرین: نیکی سپاسیان

مهلت ارسال امتیازی: ۱۱ آبان | **تمرین دوم** | مهلت ارسال نهایی: ۱۸ آبان

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روزهای مشخص‌شده است.
- در طول ترم، برای هر تمرین می‌توانید تا ۵ روز تأخیر مجاز داشته باشید و در مجموع حداکثر ۱۵ روز تأخیر مجاز خواهید داشت. توجه داشته باشید که تأخیر در تمرین‌های عملی و تئوری به صورت جداگانه محاسبه می‌شود و مجموع تأخیر هر دو نباید بیشتر از ۱۵ روز شود. پس از اتمام زمان مجاز، دو روز اضافی برای آپلود غیرمجاز در نظر گرفته شده است که در این بازه به ازای هر ساعت تأخیر، ۲ درصد از نمره تمرین کسر خواهد شد.
- اگر بخش عملی یا تئوری تمرین را قبل از مهلت ارسال امتیازی آپلود کنید، ۲۰ درصد نمره اضافی به آن بخش تعلق خواهد گرفت و پس از آن، ویدئویی تحت عنوان راهنمایی برای حل تمرین منتشر خواهد شد.
- حتماً تمرین‌ها را بر اساس موارد ذکرشده در صورت سوالات حل کنید. در صورت وجود هرگونه ابهام، آن را در صفحه تمرین در سایت کوئرا مطرح کنید و به پاسخ‌هایی که از سوی دستیار آموزشی مربوطه ارائه می‌شود، توجه کنید.
- در صورت هم‌فکری و یا استفاده از هر منابع خارج درسی، نام هم‌فکران و آدرس منابع مورد استفاده برای حل سوال مورد نظر را ذکر کنید.
- فایل پاسخ‌های سوالات نظری را در قالب یک فایل *pdf* به فرمت $HW۲\_T\_[STD\_ID].pdf$ آماده کنید و برای سوالات عملی، هریک را در یک فایل *zip* جداگانه قرار دهید و فایل *zip* اول را به فرمت $HW۲\_P۱\_[STD\_ID].zip$ و فایل *zip* دوم را به فرمت $HW۲\_P۲\_[STD\_ID].zip$ نامگذاری کرده و هرکدام را به صورت جداگانه آپلود کنید.
- گردآورندگان تمرین: محمدپارسا بشری، عرفان جعفری، مهدی طباطبایی، فاطمه السادات موسوی، محمد مولوی

## سوالات نظری (۱۰۰ نمره)

۱. (۲۰ نمره) در یک مسئله‌ی طبقه‌بندی دو کلاسه با دو ویژگی، از هر کلاس دو داده داریم:

$$\omega_۱ : \begin{pmatrix}۴\\۱\end{pmatrix}, \begin{pmatrix}۱\\۰\end{pmatrix}, \omega_۲ : \begin{pmatrix}۲\\۵\end{pmatrix}, \begin{pmatrix}۳\\۴\end{pmatrix}$$

الف) با محاسبه‌ی بردار میانگین و ماتریس کوواریانس، داده‌ها را با استفاده از PCA به فضای یک بعدی برده و تمایزپذیری کلاس‌ها را بررسی کنید. برای هر دو راستا مسئله را حل کنید.

ب) در قسمت الف یک بردار پیشنهاد دهید که اگر به همه‌ی داده‌ها اضافه شود مولفه‌ی اساسی اول تغییر نکند.

۲. (۲۰ نمره) قصد داریم مدلی بسازیم که با گرفتن ورودی $X$، متغیر خروجی $Y$ را تخمین بزنند. فرض کنید $M$ مدل ضعیف به نام‌های $f_۱(X), ..., f_M(X)$ داریم که روی نمونه‌های bootstrap شده‌ای از دیتاست اصلی آموزش داده شده‌اند و دارای بایاس و واریانس یکسان هستند. مدل نهایی به شکل زیر تعریف می‌شود:

$$f_{\text{ensemble}}(X) = \frac{۱}{M}\sum_{i=۱}^{M} f_i(X)$$

الف) بایاس و واریانس مدل $f_{\text{ensemble}}(X)$ را بر حسب بایاس و واریانس مدل‌های ضعیف محاسبه کنید. در این قسمت فرض کنید مدل‌های ضعیف از یکدیگر مستقل هستند. تحلیل کنید که افزایش تعداد مدل‌های ضعیف ($M$) چه تاثیری روی بایاس و واریانس مدل نهایی دارد.

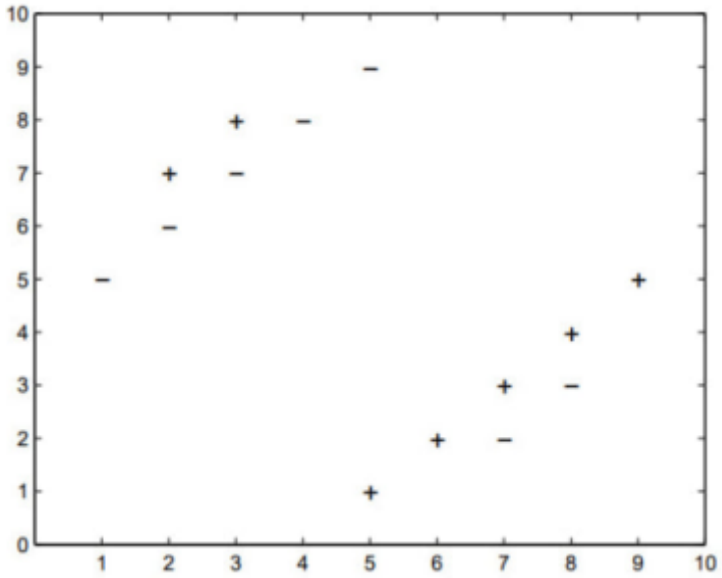
$$Var\left(\sum_{i=1}^{N} a_i X_i\right) = \sum_{i,j=1}^{N} a_i a_j Cov(X_i, X_j)$$

ب) حال فرض کنید مدل‌های ضعیف مستقل نیستند و correlation بین هر دو مدل $f_i(X)$ و $f_j(X)$ $(i \neq j)$ برابر $\rho$ است. حال مانند قسمت الف، بایاس و واریانس مدل نهایی را محاسبه کرده و تاثیر تعداد مدل‌ها $(M)$ و وابستگی بین آن‌ها $(\rho)$ را روی این مقادیر بررسی کنید.

پ) به سوالات زیر به طور خلاصه پاسخ دهید.

- آیا یادگیرنده‌های ضعیف در adaboost نیاز به مشتق‌پذیر بودن دارند؟ چرا؟
- بین boosting و bagging ، کدام یک از نظر محاسباتی گران‌تر می‌باشد؟ چرا؟

۳. (۲۰ نمره) در این سوال یک دسته‌بند KNN با متریک فاصله $L_2$ در نظر بگیرید. کلاس‌ها را تماما دوحالته (+/- ) در نظر خواهیم گرفت. به سوالات زیر با توجه به مجموعه‌دادهٔ مشخص شده در تصویر پاسخ دهید.



هر نقطه همسایه خودش هم هست

باید فرض کنیم مثلا علامت یک داده در ترین رو نمی دونیم و بعد با استفاده از همسایگی ها علامتش رو تشخیص بدیم

تعریف خطا در KNN
نسبت تعداد دسته بندی های نادرست به کل دسته بندی ها

الف) به ازای چه مقدار k خطای این دسته بند کمینه می‌شود؟ مقدار این خطا چقدر است؟
ب) چرا استفاده از مقادیر بسیار زیاد یا بسیار کم برای k می‌تواند منجر به خطا شود؟
پ) فرض کنید از روش Leave One Out Cross Validation استفاده کنیم. به ازای چه مقدار k خطای دسته‌بندی کمینه می‌شود؟ مقدار این خطا چقدر است؟(استفاده از ابزارهای sklearn برای به دست آوردن k بهینه مجاز است.).
ت) مرز تصمیم برای دسته‌بند 1-NN را برای این مجموعه‌داده در تصویر نشان دهید.

Leave-one-out cross validation is K-fold cross validation taken to its logical extreme, with K equal to N, the number of data points in the set. That means that N separate times, the function approximator is trained on all the data except for one point and a prediction is made for that point.

۴. (۲۰ نمره)

تابع هزینه الگوریتم خوشه‌بندی k-means به شکل زیر تعریف می‌شود:

$$L = \sum_{j=1}^{k} \sum_{x_i \in S_j} \|x_i - \mu_j\|^2$$

که در آن $x_1, x_2, \ldots, x_n$ نمونه‌ها و $\mu_1, \mu_2, \ldots, \mu_k$ مراکز خوشه‌ها هستند. منظور از $S_j$ نیز مجموعه‌ای از نمونه‌هاست که به مرکز $\mu_j$ نزدیک‌تر از هر خوشه‌ی دیگر هستند.

الف) یک مرحله از الگوریتم را در نظر بگیرید که برچسب داده‌ها $y_j$ ثابت است و مراکز خوشه‌ها $\mu_i$ به‌روزرسانی می‌شوند. نشان دهید که برای کمینه کردن تابع هزینه در این مرحله، کافی است میانگین هر خوشه به‌عنوان مرکز آن خوشه قرار گیرد.

مشتق بگیر صفر بذار

بله

ب) آیا الگوریتم k-means نسبت به مقداردهی اولیه مراکز خوشه‌ها حساس است؟ آیا این الگوریتم به طور تضمینی همگرا می‌شود؟

پ) در مرحله‌ای از الگوریتم k-means که در آن میانگین خوشه‌ها $\mu_i$ ثابت‌اند و برچسب‌های نقاط داده $y_j$ به‌روزرسانی می‌شوند، گاهی ممکن است یک نقطه $X_j$ به چندین مرکز خوشه با فاصله مساوی نزدیک باشد. اگر یکی از گزینه‌ها این باشد که $X_j$ در همان خوشه‌ای که در تکرار قبلی بوده باقی بماند، توضیح دهید چرا بهتر است این گزینه انتخاب شود؟ اگر این اصل رعایت نشود، چه مشکلی ممکن است پیش بیاید؟

یکی از راه هایی که برای پایدار کردن الگوریتم داشتیم این بود که اگر بعد از مدتی تغییری در مرکز خوشه ها ایجاد نشد الگوریتم را متوقف کنیم بنابراین بهتره توی همون خوشه قبلی باقی بمونه

۵. (۲۰ نمره) فرض کنید مجموعه‌ای از نقاط $x^{(1)}, \ldots, x^{(m)}$ داده شده‌اند. فرض کنید داده‌ها نرمال شده‌اند و دارای میانگین صفر و واریانس یک در هر بعد هستند. همچنین فرض کنید $f_u(x)$ تصویر نقطه $x$ در جهت بردار یکه $u$ باشد. به عبارتی اگر داشته باشیم:

مشتق خطا رو برابر صفر میذاریم
آ بر حسب یو و ایکس به دست میاد
اونا رو تو رابطه اصلی جایگذاری میکنیم
یکم ساده سازی
روی معادله به دست آمده روش لاگرانژ رو اجرا میکنیم

$$V = \{au : a \in \mathbb{R}\}$$

آنگاه:

$$f_u(x) = \arg\min_{v \in V} \|x - v\|^2$$

نشان دهید بردار $u$ که خطای MSE بین نقاط و تصویر آن‌ها را کمینه می‌کند، همان مولفه اصلی اول ( $PC_1$ ) است. به عبارتی نشان دهید که:

$$\arg\min_{u:u^\top u=1} \frac{1}{m} \sum_{i=1}^{m} \left\|x^{(i)} - f_u(x^{(i)})\right\|^2$$

برابر اولین مولفه اصلی است.

---

## سوالات عملی (۱۰۰ نمره)

۱. (۱۰۰ نمره)

(آ) (۵۰ نمره) برای پاسخ به بخش اول تمرین عملی ابتدا نوتبوک KNN-Ensemble را باز کنید و سپس مراحل را مطابق آن‌چه از شما خواسته شده انجام دهید. در نهایت، مقادیر پیشبینی شده برای دیتاست test.csv را مطابق آن چه در نوتبوک توضیح داده شده است در یک فایل با نام result.csv که شامل یک ستون با نام target میباشد، ذخیره کنید. فایل خروجی و فایل نوتبوک را در یک فایل zip قرار دهید و آن را به فرمت HW2_P1_[STD_ID].zip نامگذاری کرده و آپلود کنید.

توجه بفرمایید که این سوال دارای داوری خودکار میباشد و ۱۵ نمره از ۵۰ نمره به این بخش تعلق دارد.

(ب) (۵۰ نمره) برای پاسخ به بخش دوم تمرین عملی دوم تنها کافی است نوتبوک Clustering-PCA را تکمیل کرده و مطابق فرمت HW2_P2_[STD_ID].zip نامگذاری کرده و آپلود کنید.