**به نام خدا**



**دانشگاه صنعتی امیرکبیر**

**دانشکده‌ی مهندسی کامپیوتر**

# تمرین سری دوم یادگیری ماشین

**دکتر احسان ناظرفرد**

طراح سوال:

محمدرضا امامی ناصری

سید اردلان قریشی

پاییز ۱۳۹۹

**توضیحات مهم:**

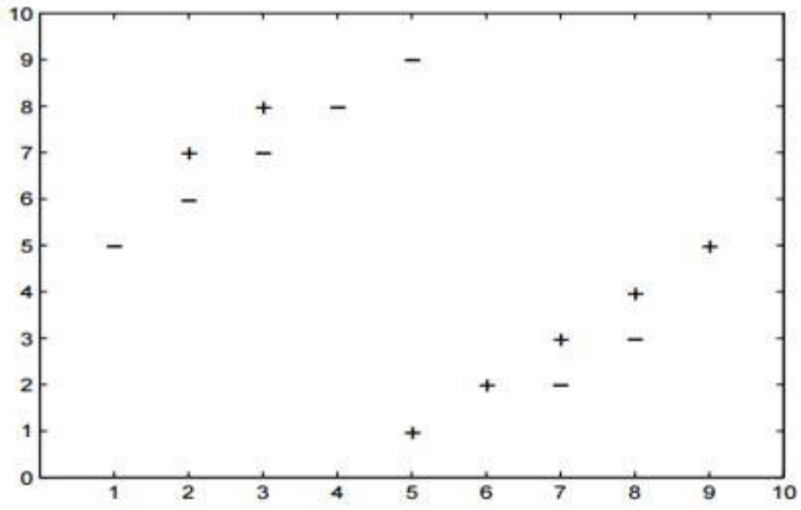
- تمامی مستندات خود شامل گزارش و کدهای خود را در یک فایل فشرده با فرمت zip ذخیره کرده و با عنوان #StudentId_HW2.zip بارگذاری نمایید (به عنوان مثال 99131000_HW2.zip).
- مهلت انجام تمرین تا **ساعت ۲۳:۵۵ روز چهارشنبه مورخ ۲۶ آذر** می‌باشد و به هیچ وجه تمدید نمی‌شود.
- تمرین بدون گزارش فاقد ارزش می‌باشد و **نمره‌ای به آن تعلق نمی‌یابد**.
- تا حد ممکن سعی کنید اصول لازم برای گزارش مهندسی را رعایت نمایید (به بهترین گزارش نمره تشویقی تعلق می‌گیرد).
- مطابق قوانین دانشگاه هرگونه کپی برداری **ممنوع** می‌باشد و در صورت مشاهده نمره‌ی **هر دو طرف** صفر در نظر گرفته می‌شود.
- شما مجاز هستید برای تمامی تمرین‌ها **۷ روز در کل** و **با سقف حداکثر ۳ روز برای هر تمرین**، تاخیر بدون کسر نمره داشته باشید. به ازای هر روز تاخیر بیشتر، ۱۰٪ از نمره‌ی تمرین مربوطه کسر می‌شود.
- در صورت داشتن هرگونه ابهام می‌توانید از طریق ایمیل زیر سوال خود را مطرح نمایید:

MLAUTFALL99@gmail.com

## سوالات تشریحی

**۱**- برای یافتن بهترین مقدار پارامتر K در الگوریتم K-نزدیک‌ترین همسایه[1]، چه راهکاری را پیشنهاد می‌کنید؟ **۳٪**

**۲**- با توجه به شکل زیر به سوالات پاسخ دهید. **۸٪**




الف) بهترین مقدار K برای الگوریتم K-نزدیک‌ترین همسایه زمانی که از روش LOOCV[2] استفاده می‌شود را محاسبه کنید. دقت[3] الگوریتم را به ازای این K گزارش نمایید.

ب) مشکل انتخاب مقدار K خیلی بزرگ و خیلی کوچک چیست؟ توضیح دهید.

ج) نقطه (2,1) با توجه به K به دست آمده به کدام کلاس تعلق می‌یابد؟

**۳**- هر کدام از الگوریتم‌های K-نزدیک‌ترین همسایه و درخت تصمیم[4] را از نظر پارامتریک و غیر پارامتریک[5] بودن بررسی کنید. **۲٪**

**۴**- هر کدام از الگوریتم‌های K-نزدیک‌ترین همسایه و درخت تصمیم را از Generative و Discriminative بودن بررسی کنید. **۲٪**

[1] K-Nearest Neighbors algorithm (KNN)
[2] Leave One Out Cross Validation
[3] accuracy
[4] decision tree
[5] parametric and nonparametric

**۵**- به چه الگوریتم‌هایی تنبل[6] گفته می‌شود؟ K-نزدیک‌ترین همسایه تنبل است یا خیر؟ توضیح دهید. **۲٪**

**۶**- هرس[7] درخت تصمیم چه تاثیری بر بیش برازش[8] دارد؟ این هرس چه زمانی باید انجام شود؟ توضیح دهید.**۴٪**

**۷**- در جدول ۱ مجموعه‌داده‌ای نمایش داده شده است که در آن افراد با توجه به ویژگی‌هایی مثل سن، درآمد و... اقدام به خرید یا عدم خرید یک کالا کرده‌اند. هدف ما تخمین[9] این است که آیا فرد مورد نظر قصد خرید کالا را دارد یا خیر. **۱۰٪**

| age | income | student | credit | Buy |
|---|---|---|---|---|
| youth | high | no | fair | - |
| youth | high | no | excellent | - |
| middle | high | no | fair | + |
| senior | medium | no | fair | + |
| senior | low | yes | fair | + |
| senior | low | yes | excellent | - |
| middle | low | yes | excellent | + |
| youth | medium | no | fair | - |
| youth | low | yes | fair | + |
| senior | medium | yes | fair | + |
| youth | medium | yes | excellent | + |
| middle | medium | no | excellent | + |
| middle | high | yes | fair | + |
| senior | medium | no | excellent | - |

جدول ۱

الف- با توجه به ویژگی آنتروپی[10] و بهره اطلاعات[11] درخت تصمیم بهینه[12] را برای این مجموعه داده بیابید.

ب- داده‌های زیر را با توجه به درخت تصمیم به دست آمده، دسته‌بندی[13] کنید.

$X_1$ = (age = youth, income = high, student = yes, credit = fair)

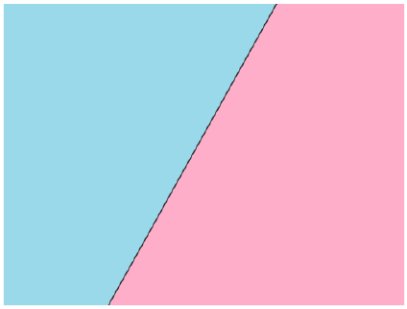
$X_2$ = (age = senior, income = low, student = no, credit = excellent)

$X_3$ = (age = middle-aged, income = medium, student = no, credit = fair)

---

[6] lazy
[7] pruning
[8] overfitting
[9] predict
[10] entropy
[11] information gain
[12] optimal
[13] classify

**۸**- در دسته‌بندی[۱۴] داده‌ها با درخت تصمیم می‌توانیم با افزایش عمقِ درخت[۱۵] توابع بسیار پیچیده‌ای بسازیم. آیا برای مسائل جداپذیرِ خطی[۱۶] مانند شکل زیر می‌توان دسته‌بندی را با عمق محدود درخت انجام داد؟ **۴٪**



**۹**- الگوریتم جنگل تصادفی[۱۷] را مختصر توضیح دهید. چرا این الگوریتم با محدود نگه داشتن عمق درخت می‌تواند داده‌ها را جدا کند؟ **۵٪**

[14] classification
[15] tree-depth
[16] linearly separable
[17] random forest

## Weka

هدف از این بخش آشنایی با ابزار شناخته‌شده‌ی وِکا است. برای انجام تمرینات این بخش می‌توانید از آموزش‌های موجود در اینترنت کمک بگیرید.

(به عنوان نمونه: https://www.tutorialspoint.com/weka/index.htm)

**۱**- مجموعه‌داده‌ی پیوست شده به نام breast-cancer.arff را بارگذاری کرده و با استفاده از درخت تصمیم (J48) داده‌ها را دسته‌بندی کنید. (۷۰ درصد داده‌ها را به عنوان داده‌های آموزش[18] و ۳۰ درصد داده‌ها را به عنوان داده آزمون[19] در نظر بگیرید.) **۴٪**

**خروجی مورد نظر**: شکل درخت تصمیم نهایی، ماتریس درهم ریختگی[20]، مقادیر TP، FP، FN، Precision، Recall، F1measure و تحلیل لازم.

**۲**- پارامتر unpruned چه چیزی را کنترل می‌کند؟ این مقدار را از false به true تغییر داده و دوباره موارد قسمت ۱ را انجام دهید. درخت آموزش داده شده در این قسمت چه تفاوتی با قسمت ۱ دارد؟ توضیح دهید. **۴٪**

**خروجی مورد نظر**: شکل درخت تصمیم نهایی، ماتریس درهم ریختگی، مقادیر TP، FP، FN، Precision، Recall، F1measure و تحلیل لازم.

**۳**- حال موارد قسمت ۱ و ۲ را بار دیگر، این بار با اعمال ۱۵ درصد نویز[21] به ریشه‌ی درخت‌های دو قسمت قبل تکرار کنید و نتایج حاصل را با نتایج مراحل قبلی مقایسه کرده و بررسی کنید که هرس کردن درخت چه تاثیری در برخورد با نویز دارد. **۲٪**

**خروجی مورد نظر**: ماتریس درهم ریختگی، مقادیر TP، FP، FN، Precision، Recall، F1measure و تحلیل لازم.

[18] train
[19] test
[20] confusion matrix
[21] noise

## سوالات پیاده‌سازی

**توضیحات مهم:**

- در روند اجرا انتخاب مقادیر برای تقسیم داده‌ها به مجموعه آموزش، ارزیابی و... به عهده دانشجو می‌باشد.
- حتما پارامترهای انتخاب شده برای برنامه خود و هرگونه شرایطی که درنظر گرفته‌اید را در گزارش خود بیاورید.
- برای بهبود سرعت برنامه توصیه می‌شود از عملیات ماتریسی استفاده کنید.
- در هر مرحله، نتایج خود را تحلیل کنید.
- کدهای خود را برای خوانایی بیشتر **کامنت گذاری** کنید.
- در تمامی سوال‌ها تنها مجاز به استفاده از کتابخانه‌هایnumpy، matplotlib و pandas می‌باشید.
- در پیاده‌سازی بخش‌های مختلف، امکان استفاده از کتابخانه‌های آماده مرتبط با الگوریتم‌های یادگیری ماشین را به طور کلی ندارید. **موارد مجاز در صورت سوال ذکر شده است.**
- گذاشتن عنوان برای نمودارها و برچسب گذاری محورهای نمودار الزامی می‌باشد.

در این بخش برای دسته‌بندی از مجموعه‌داده‌های مرتبط با بیماری سرطان پستان[۲۲] استفاده شده است. دلیل این امر اول مواجه شدن عزیزانِ دو طراح این سری تمرین با این بیماری شوم و دوم نشان دادن یکی از هزاران کاربرد الگوریتم‌های دسته‌بندی در دنیای واقعی بوده است. با توجه به اینکه این بیماری در صورت تشخیص سریع می‌تواند درمان بسیار بهتری داشته باشد، قسمتی از پایان‌نامه‌ی مادرم که خود در دوران دست و پنجه نرم کردن با این بیماری اقدام به نوشتن و دفاع از آن کرده است را به امید اطلاع‌رسانی هرچند کوچک آورده‌ام. سید اردلان قریشی

سرطان پستان اولین سرطان شایع و دومین علت مرگ ناشی از سرطان در بین زنان ۳۵ تا ۵۵ ساله ایران بوده و سالانه حدود ۷ هزار مورد جدید به بیماران قبلی اضافه می‌شود. از هر ۸ زن ایرانی در فاصله سنی ۴۵ تا ۵۵ سال، یک نفر شانس ابتلا به این سرطان را دارد، در ایران سن ابتلا در زنان در سال‌های اخیر کاهش یافته است، به عبارتی در دنیا مادربزرگ‌ها به این بیماری مبتلا می‌شوند و در ایران مادران!

این بیماری تاثیرات شدید روانی ایجاد می‌کند و تشخیص و درمان آن، تجربه‌ای همراه با استرس و اضطراب می‌باشد. با اینکه جراحی رایج‌ترین درمان سرطان می‌باشد، اما این بیماران پس از جراحی از یک سو با مشکلاتی از جانب سرطان و جراحی و از سوی دیگر با درمان‌هایی مثل شیمی‌درمانی و عوارض جانبی ناخوشایندی مثل ریزش مو، حالت تهوع، اِدِم لنفاوی[۲۳] و سایر مشکلات مواجه می‌شوند. درمان‌های طولانی، توانایی زنان را دربرقراری نقش اجتماعی به عنوان زن خانه‌دار یا شاغل، مورد تهدید قرار می‌دهند. سطح بالای استرس، تأثیر منفی طولانی مدت بر خود باوری زنان دارد که در نهایت تأثیر نامطلوبی بر عملکرد خانوادگی و نقش زناشویی گذاشته و همچنین منجر به پایین آمدن سطح کیفی زندگی آنان می‌شود. (محمدی گل، ۱۳۹۵).

توصیه می‌شود که معاینه‌ی شخصی و چک کردن علایم از سن ۲۰ سالگی به صورت ماهانه توسط هر فرد انجام شود. علاوه بر آن از سن ۲۰ تا ۳۹ سالگی هر ۳ سال این معاینه توسط پزشک نیز انجام گردد و از ۴۰ سالگی به بعد به طور سالانه معاینه توسط پزشک انجام شود. (همایی فاطمه، ۱۳۹۰).

[22] breast cancer

[23] lymphedema

**۱**- یک تابع[۲۴] بنویسید که با گرفتن ورودی[۲۵]های مجموعه داده، K و معیار فاصله[۲۶]، الگوریتم KNN را اجرا کند. از آن تابع برای دسته‌بندی مجموعه‌داده‌ی mammographic_masses.data استفاده کنید. **۲۵٪**

- برای ارزیابی[۲۷] استفاده از 10-fold cross validation الزامی است.
- در این مجموعه داده برخی نمونه[۲۸]ها دارای ویژگی‌هایی با مقادیر نامعلوم[۲۹] می‌باشند. به ابتکار خود روشی را برای حل این مشکل بیابید و روش خود را در گزارش شرح دهید.
- پیاده‌سازی ماتریس درهم ریختگی و k-fold cross validation الزامی نیست و می‌توانید از کتابخانه‌ی آماده برای این دو مورد استفاد کنید. پیاده‌سازی این موارد توسط خود دانشجو نمره تشویقی دارد. (در گزارش ذکر شود.)
- پیش پردازش[۳۰] لازم را بر روی مجموعه داده انجام داده و در گزارش خود بیاورید.

الف) الگوریتم KNN را به ازای مقادیر مختلف ۱،۳،۵،۷،۱۵،۳۰ برای K و با فاصله اقلیدسی[۳۱] اجرا کرده و تاثیر مقادیر مختلف K را تحلیل کنید.

**خروجی مورد نظر**: دقت الگوریتم و ماتریس درهم ریختگی

ب) به ازای بهترین مقدار K که در قسمت الف یافته‌اید و با فاصله‌های اقلیدسی، منهتن[۳۲] و کسینوسی[۳۳]، الگوریتم را اجرا کنید.

**خروجی مورد نظر**: دقت الگوریتم و ماتریس درهم ریختگی

ج) با استفاده از کتابخانه‌های آماده و به ازای مقادیر مختلف ۱،۳،۵،۷،۱۵،۳۰ برای K و فاصله اقلیدسی، داده‌ها را دسته‌بندی کرده و پیاده‌سازی خود را از نظر دقت و سرعت با این کتابخانه مقایسه کنید.

**خروجی مورد نظر**: دقت الگوریتم، ماتریس درهم ریختگی و زمان اجرای هر الگوریتم

**۲**- در این قسمت از الگوریتم KNN برای رگرسیون استفاده نمایید. مجموعه داده‌ی regression.xlsx را با نسبت ۷۰ به ۳۰ تقسیم کرده و سپس بهترین مقدار برای K را با آزمون و خطا بیابید. خطای MSE را برای این مدل برای هر دو مجموعه‌ی داده آزمون و آموزش گزارش کنید. **۱۵٪**

**خروجی مورد نظر**: خطای MSE برای هر دو مجموعه آزمون و آموزش

[24] function
[25] input
[26] distance criterion
[27] evaluation
[28] instance
[29] missing value
[30] preprocessing
[31] Euclidean distance
[32] Manhattan distance
[33] cosine distance/similarity

**۳**- درخت تصمیم بهینه را با استفاده از کتابخانه‌های آماده برای مجموعه داده‌ی breast-cancer-wisconsin-train.data آموزش داده و مجموعه داده‌ی breast-cancer-wisconsin-test.data را دسته‌بندی کنید. **۱۰٪**

- در این مجموعه داده برخی نمونه‌ها دارای ویژگی‌هایی با مقادیر نامعلوم می‌باشند. به ابتکار خود روشی را برای حل این مشکل بیابید و روش خود را در گزارش شرح دهید.
- پیش پردازش‌های لازم را بر روی مجموعه داده انجام داده و در گزارش خود بیاورید.

**خروجی مورد نظر**: دقت الگوریتم برای مجموعه داده آزمون و ماتریس درهم ریختگی

با آرزوی موفقیت!