توضیحات مهم:

- تمامی مستندات خود شامل گزارش و کدهای خود را در یک فایل فشرده با فرمت zip ذخیره کرده و با عنوان studentID_HW1.zip بارگذاری نمایید(بعنوان مثال 99131000_HW1.zip).
- مهلت انجام تمرین تا ساعت ۲۳:۵۵ تاریخ ۱۴۰۰/۰۱/۱۵ می‌باشد و **به هیچ وجه تمدید نمی‌شود.**
- استفاده از کتابخانه‌های رایج در یادگیری ماشین بلا مانع است.
- برای سهولت در انجام تمرین می‌توانید از پلتفرم کولب گوگل استفاده نمایید.
- ملاک اصلی انجام تمرین گزارش آن است و ارسال کد بدون گزارش فاقد ارزش است. یک فایل pdf تهیه کرده و برای هر سوال، ورودی، خروجی و توضیحات مربوطه را بصورت جامع گزارش کنید.
- تا حد ممکن سعی کنید اصول لازم برای گزارش مهندسی را رعایت نمایید. (به بهترین گزارش نمره‌ی تشویقی تعلق می‌گیرد.)
- مطابق قوانین دانشگاه هرگونه کپی‌برداری ممنوع می‌باشد و در صورت مشاهده نمره‌ی طرفین **صفر** 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 هر گونه ابهام می‌توانید از طریق ایمیل زیر سوال خود را مطرح نمایید:

Z.Khalvandi77@gmail.com
Najmeh.Mohammadbagheri77@gmail.com

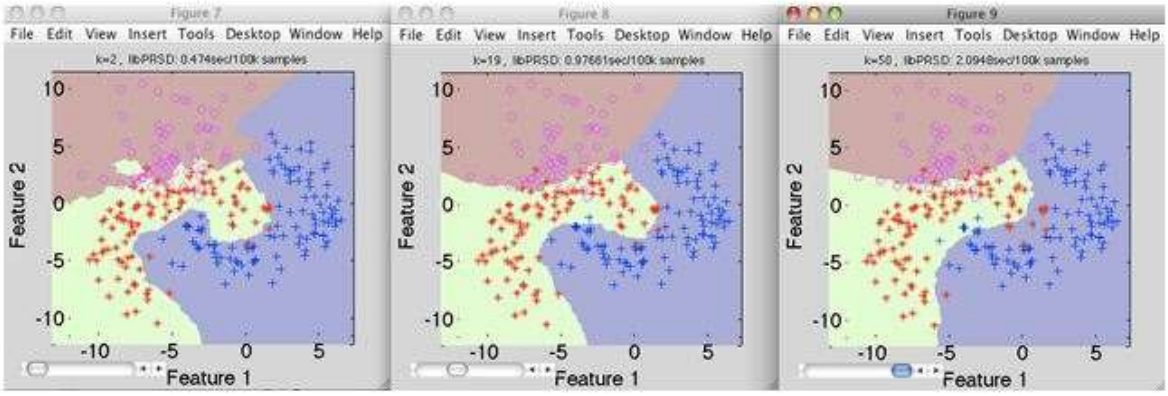
بخش اول: سوالات تشریحی

۱- هوش مصنوعی و یادگیری ماشین را تعریف کنید و ارتباط بین این دو مفهوم را شرح دهید.

۲- دو مفهوم همبستگی و کواریانس چه تفاوتی دارند؟

۳- مدل‌های پارامتری و غیرپارامتری را تعریف و سپس مشخص کنید K – نزدیکترین همسایه، درخت تصمیم و رگرسیون خطی در کدام دسته قرار می‌گیرند.

۴- شکل ۱ نتیجه‌ی اجرای الگوریتم نزدیکترین همسایه را با سه مقدار k متفاوت بر روی یک مجموعه داده نشان می‌دهند، k انتخاب شده در این سه حالت را با هم مقایسه کنید.



شکل ۱

۵– هرس درخت تصمیم چه تاثیری بر بیش برازش دارد؟ این هرس چه زمانی باید انجام شود؟ توضیح دهید.

۶– در جدول ۱ مجموعه دادهای نمایش داده شده است که در آن داروی مورد نیاز بیماران با توجه به ویژگی‌هایی مثل سن، جنسیت، سطح فشارخون و سطح کلسترول خون مشخص می‌شود.

Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	Hiigh	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A
p15	Middle-age	F	Low	Normal	?

الف) با توجه به ویژگی آنتروپی و بهره اطلاعات درخت تصمیم بهینه را برای این مجموعه داده بیابید.
ب) به کمک درخت بدست آمده، نوع دارو را برای بیماری که مشخصاتش در سطر آخر آمده است مشخص کنید.

بخش دوم: آشنایی با نرم‌افزار weka

نرم‌افزار وکا یک مجموعه از الگوریتم‌های یادگیری ماشین است که در حوزه‌ی داده‌کاوی استفاده می‌شوند. شامل ابزاری برای آماده‌سازی داده‌ها، حل مسائل دسته‌بندی، رگرسیون، خوشه‌بندی و مصورسازی داده‌ها است. هدف از این قسمت تمرین آشنایی با این نرم‌افزار است.

برای انجام این بخش می‌توانید از محتواهای موجود در وب کمک بگیرید.[1]

مجموعه داده‌ی diabetes.arff را در وِکا بارگذاری کنید و بخش‌های الف تا ج را انجام دهید.

[1] https://www.tutorialspoint.com/weka/index.htm

الف) داده‌ها را با درخت تصمیم (J48) دسته‌بندی کنید. (داده‌ها را به دو بخش ۷۰٪ آموزش و ۳۰٪ آزمون تقسیم کنید.)

ب) نقش پارامتر unpruned را شرح دهید و مقدار آن را از حالت پیش‌فرض به حالت True تغییر دهید و آزمایش بخش قبل را تکرار کنید و نتایج جدید را گزارش کنید و با بخش قبل مقایسه کنید.

ج) بخش‌های الف و ب را با افزودن ۱۵ درصد نویز به ریشه‌ی درخت تکرار کنید. تاثیر نویز را بر دسته‌بندی بررسی کنید. نتیجه اعمال هرس برای مقابله با نویز را بررسی کنید.

توجه:

۱. در تمام بخش‌ها درخت بدست آمده، مقادیر precision ، recall ، Accuracy و ماتریس درهم ریختگی را به همراه تحلیل نتایج گزارش کنید.

بخش سوم: پیاده‌سازی

رگرسیون

۱- برای حل بخش‌های زیر مجموعه داده‌ی insurance.csv که به فایل تکلیف پیوست شده‌است را بارگذاری نمایید. در این مجموعه داده مشخص شده‌است که شرکت بیمه با توجه به مشخصات هر بیمار چه هزینه‌ای را پرداخت کرده است. این مشخصات شامل سن، جنسیت، تعداد فرزندان، مصرف یا عدم مصرف دخانیات و محل سکونت می‌باشد.

الف) در مجموعه داده ذکر شده، برخی از ویژگی‌ها بصورت غیر عددی هستند. برای استفاده از داده‌ها در رگرسیون باید این ویژگی‌ها به حالت عددی تبدیل شوند. بدین منظور پیش پردازش‌های لازم را انجام داده و مراحل کار را در گزارش ذکر کنید.

ب) ابتدا قصد داریم با استفاده از رگرسیون خطی تک متغیره، هزینه‌ی بیمه را تخمین بزنیم. بدین منظور باید هر بار یکی از ویژگی‌ها را انتخاب کنید و مدل رگرسیون را آموزش دهید و در نهایت متغیر مناسب را انتخاب کنید.

در این بخش مقادیر خطای آموزش متغیرها را با هم مقایسه و شیب خط و عرض از مبدا بهترین مدل را گزارش کنید.

ج) در این بخش با هدف افزایش دقت مدل، از تمام ویژگی‌های ورودی استفاده کنید و مدل رگسیون چند متغیره را آموزش دهید. نتایج این بخش را با بخش الف مقایسه و تحلیل کنید.

۲- الف) مجموعه داده‌ی Dataset1.csv را رسم کنید.

ب) مدل‌های رگرسیون خطی، درجه دو، درجه سه و چهار را بر روی داده‌ها اجرا کنید. منحنی بدست آمده برای هر حالت را در کنار هم و بر روی داده‌ها رسم کنید و خطای هر مدل را گزارش کنید و بررسی کنید که کدام مدل رگرسیون برای تخمین داده‌ها مناسب است؟

پ) در این بخش از مدل KNN برای رگرسیون بر روی داده‌های bmd.csv استفاده کنید. داده‌ها را به نسبت مناسب به دو بخش آزمون و آموزش تقسیم کنید و سپس حداقل سه مقدار متفاوت برای K را آزمایش کنید و برای هر حالت مقدار خطای آموزش و آزمون را گزارش و تحلیل کنید.

K–نزدیکترین همسایه

۱- در این مساله می‌خواهیم با استفاده از الگوریتم k-نزدیک‌ترین همسایه و درخت تصمیم برای دسته‌بندی استفاده کنیم. مجموعه داده bdiag.csv شامل چندین جزئیات تصویربرداری از بیمارانی بود که برای آزمایش سرطان سینه بیوپسی شده بودند .تشخیص متغیر بافت بیوپسی شده را به عنوان M = بدخیم یا B = خوش خیم طبقه بندی می کند.

الف) یک تابع بنویسید که با گرفتن ورودی‌های مجموعه‌داده، معیار فاصله و K، الگوریتم k-نزدیک‌ترین همسایه را اجرا کند. الگوریتم را به ازای مقادیر ۱، ۳، ۷، ۱۵ و ۳۰ برای K و با فاصله‌ی اقلیدسی اجرا کرده و نتایج دسته‌بندی را گزارش کنید. در نهایت بهترین K را انتخاب و معیارهای ارزیابی را بر روی داده‌ها آزمون بدست آورید.

توجه:

۱. اعمال پیش‌پردازش‌های لازم ضروری است. همچنین برای ارزیابی استفاده از 10-fold cross validation الزامی است.

۲. داده‌ها را به سه بخش آموزش، آزمون و ارزیابی تقسیم کنید.

۳. برای انجام این بخش می‌توانید از کلاس KNeighborsClassifier از کتابخانه‌ی sklearn استفاده کنید.

درخت تصمیم

۱- در این بخش باید به کمک مدل درخت تصمیم مجموعه داده‌یAgaricus-lepiota.data را دسته‌بندی کنید. بدین منظور حداقل سه مقدار مختلف حداکثر عمق را برای درخت بیازمایید و عمق بهینه برای دسته‌بندی این داده‌ها پیدا کنید.

ماتریس درهم ریختگی، دقت، precision ، recall را برای هر حالت گزارش کنید.

توجه:

۱. برای انجام این بخش می‌توانید از کلاس DecisionTreeClassifier از ماژول sklearn استفاده کنید.