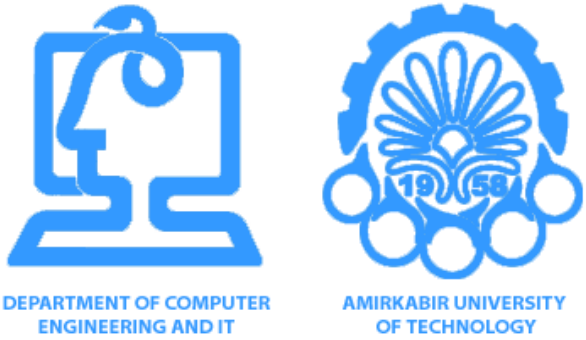

دانشگاه صنعتی امیرکبیر
دانشکده مهندسی کامپیوتر

پروژه نهایی درس یادگیری ماشین

دکتر ناظرفرد

پاییز ۹۸

- تمامی مستندات خود شامل گزارش و کدهای خود را در یک فایل فشرده با فرمت زیپ ذخیره کرده و با عنوان #studentId_HW.zip بارگزاری نمایید.

- مهلت انجام این تمرین تا ساعت **۲۳:۵۵ روز ۱۸ بهمن** می‌باشد **و با توجه به زمان قفل نمرات هیچ وجه تمدید نمی‌شود**.
- تمرین بدون گزارش فاقد ارزش می‌باشد و نمره‌ای به آن تعلق **نمی‌یابد**.
- کامنت گذاری کدها در حد **لازم و کافی** الزامی می‌باشد.
- گذاشتن عنوان[1] برای نمودارها و برچسب‌گذاری[2] محورهای نمودار **الزامی** می‌باشد.
- مطابق قوانین دانشگاه هرگونه کپی‌برداری ممنوع می‌باشد و در صورت مشاهده نمره هر **دو طرف صفر** در نظر گرفته می‌شود.

در این پروژه تعدادی مجموعه داده برای انواع مسائل مختلف (رگرسیون[3]، کلاس‌بندی[4] و خوشه‌بندی[5]) مشخص شده است که هر کدام دارای چالش‌های منحصر به فردی می‌باشند. هدف از این پروژه آشنایی هرچه بیشتر شما با مشکلاتی می‌باشد که می‌تولند در چالش‌های دنیای واقعی رخ دهد. بدین منظور شما موظف می‌باشید که مدل‌های مختلفی را که در این درس یادگرفته‌اید بر روی هر کدام از مسائل آزمایش نمایید و بهترین مدل را بیابید. به عنوان مثال برای مسئله کلاس‌بندی می‌توانید از الگوریتم‌هایی مانند k نزدیک‌ترین همسایه، ماشین‌های بردار پشتیبان[6]، روش‌های ترکیبی[7] و حتی شبکه‌های عصبی استفاده نمایید و دقت هر کدام را گزارش نمایید. در نهایت چیزی که انتظار می‌رود در گزارش قرار داده شود به شرح زیر است:

- معیارهای ارزیابی خواسته شده در اجرای الگوریتم‌های مختلف بر روی هر کدام از مسائل در قالب جدول و مقایسه کمی نتایج.
- تحلیل نتایج به‌دست‌آمده به نحوی که دلایل قانع کننده برای عملکرد بهتر مدل را شامل شود.

در انجام این پروژه به نکات زیر توجه نمایید:

- یکی از اهداف دیگر این پروژه آشنایی شما با کتابخانه‌های آماده‌ای است که برای این مسائل طراحی شده‌اند. به همین علت استفاده از هر نوع کتابخانه‌ای بلامانع می‌باشد.
- مقایسه کمی میان این مدل‌ها زمانی می‌تواند معتبر باشد که از روش‌های تخمین دقت یا خطا مانند بوت‌استرپ[8] یا کراس‌ولیدیشن[9] استفاده شود (لذا استفاده از این روش‌ها الزامی می‌باشد.)
- در حل این مسائل می‌توانید بسته به خلاقیت خود از تکنیک‌های مختلف یادگیری ماشین استفاده نمایید و هیچ محدودیتی در انتخاب روش وجود ندارد ولی می‌توانید با استفاده از آموخته‌های خود بهترین روش‌ها را قبل از پیاده‌سازی با توجه به ویژگی‌های مجموعه داده‌ها حدس بزنید.

---

[1] Title
[2] Labeling
[3] Regression
[4] Classification
[5] Clustering
[6] Support Vector Machines
[7] Ensemble methods
[8] Bootstarp
[9] Cross Validation

- برای مسئله کلاس‌بندی و مسئله رگرسیون بهترین مدلی را که یافته‌اید ذخیره نمایید. در روز ارائه، شما باید مدل خود را بارگزاری نمایید تا بتوانید بر روی مجموعه داده تست (که در روز ارائه داده می‌شود) امتحان نمایید و معیار ارزیابی مشخص‌شده برای هر کدام را محاسبه نمایید.
- در هر کدام از بخش‌ها به اولین نفر (بر اساس معیار ارزیابی هر کدام) ۱۰ درصد نمره تشویقی و دومین نفر ۵ درصد نمره تشویقی اختصاص داده خواهد شد.
- دوستانی که با شبکه‌های عصبی آشنایی دارند، می‌توانند از گونه‌های مختلف این شبکه‌ها نیز برای حل مسائل مذکور استفاده نمایند.
- یکی از بهترین روش‌ها برای یافتن نقاط ضعف در حل مسائل کلاس‌بندی، به‌دست‌آوردن ماتریس پریشانی نتایج آن می‌باشد. لذا با توجه به خروجی آن می‌توانید مدل خود را ارزیابی نمایید.
- با توجه به محدودیت زمان قفل نمرات، پس از تاریخ مشخص شده برای این پروژه، دیگر امکان تحویل گرفتن آن وجود نخواهد داشت.
- در مسائل زیر، الگوریتم‌های مختلف به معنی استفاده از یک الگوریتم (ماشین بردار پشتیبان) با حالت‌های مختلف (کرنل) نیست، بلکه منظور استفاده از الگوریتم‌های مختلف همانند k نزدیک‌ترین همسایه، درخت تصمیم و ... می‌باشد ولی وقتی از شما مدل‌های مختلفی خواسته شده است، مجاز می‌باشید یک الگوریتم با گونه‌های مختلف را استفاده نمایید.
- ارائه این پروژه بعد از ۱۸ بهمن خواهد بود که اطلاع‌رسانی خواهد شد.
- دوستانی که هنوز در سامانه edmodo ثبت‌نام نکرده‌اند، هرچه سریع‌تر ثبت‌نام کرده و تمرینات خود را بارگزاری نمایند (کد ثبت‌نام درس در این سامانه fuhpvd می‌باشد).
- در صورت داشتن هر گونه سوال می‌توانید سوالات خود را از طریق ایمیل زیر مطرح نمایید:

Machinelearningf19@gmail.com

۱. مسئله رگرسیون:

بر روی مجموعه داده مشخص شده ، حداقل ۴ مدل مختلف را آزمایش نمایید و بر اساس خطای میانگین مربعات بهترین مدل رابیابید. در شکل زیر ویژگی‌های مختلف این مجموعه داده نشان داده شده است. شما می‌توانید به صورت شهودی و یا با استفاده از تکنیک‌های استخراج ویژگی بهترین ویژگی‌ها را بیابید.

توضیحات در رابطه با مجموعه داده: این مجموعه داده شامل اطلاعات مربوط به اجاره دوچرخه در سال‌های ۲۰۱۱ و ۲۰۱۲ در سامانه capital bikeshare می‌باشد که در برخی از شهرهای آمریکا مورد استفاده قرار می‌گیرد. تعداد دوچرخه‌های اجاره شده به شدت به شرایط آب و هوایی و برخی ویژگی‌های دیگر که در زیر بیان شده است بستگی دارد. ویژگی‌هایی که در این مجموعه داده وجود دارد در شکل زیر نشان داده شده است. بر این اساس شما باید تعداد دوچرخه‌های اجاره شده را با توجه به سایر ویژگی‌ها تخمین بزنید. این مجموعه داده دو حالت دارد که شما مختار به استفاده از هر کدام از آن‌ها می‌باشید. در حالت اول که با عنوان "day.csv" مشخص شده است، تعداد دوچرخه‌ها در روزهای پیاپی گردآوری شده‌اند و بررسی در سطح روزانه می‌باشد. این حالت دارای ۷۳۱ رکورد می‌باشد. در حالت دوم تغداد دوچرخه‌ها در ساعات پیاپی مشخص شده‌اند و دارای ۱۷۳۷۹ رکورد می‌باشد.

Data Dictionary

Column Position	Atrribute Name	Definition	Data Type	Example	% Null Ratios
1	instant	Record Index	Quantitative	190, 7, 17180	0
2	dteday	Date (Format: YYYY-MM-DD)	Quantitative	2012-12-23, 2012-01-01, 2012-06-24	0
3	season	Season (1: springer, 2: summer, 3: fall, 4: winter)	Quantitative	1, 2, 4	0
4	yr	Year (0: 2011, 1:2012)	Quantitative	0, 1	0
5	mnth	Month (1 to 12)	Quantitative	1, 6, 12	0
6	hr	Hour (0 to 23) - Not in day.csv dataset	Quantitative	4, 6, 14	0
7	holiday	Weather day is holiday or not	Quantitative	0, 1	0
8	weekday	Day of the week	Quantitative	0, 6, 3	0
9	workingday	Working Day: If day is neither weekend nor holiday is 1, otherwise is 0	Quantitative	0, 1	0
10	weathersit	Weather Situation (1: Clear, Few clouds, Partly cloudy, Partly cloudy; 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist; 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds, 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog)	Quantitative	1, 2, 3	0
11	temp	Normalized temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=-8, t_max=+39 (only in hourly scale)	Quantitative	0.08, 0.22, 0.34	0
12	atemp	Normalized feeling temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=-16, t_max=+50 (only in hourly scale)	Quantitative	0.0909, 0.2727, 0.303	0
13	hum	Normalized humidity. The values are divided to 100 (max)	Quantitative	0.53, 0.8, 0.31	0
14	windspeed	Normalized wind speed. The values are divided to 67 (max)	Quantitative	0.194, 0, 0.2985	0
15	casual	Count of casual users	Quantitative	0, 2, 57	0
16	registered	Count of registered users	Quantitative	1, 0, 118	0
17	cnt	Count of total rental bikes including both casual and registered	Quantitative	1, 2, 175	0

۲. مسئله کلاس‌بندی:

بر روی مجموعه داده مشخص شده زیر، حداقل ۴ الگوریتم مختلف که هرکدام شامل مدل‌هایی با پارامترهای متفاوت است آزمایش نمایید و بر اساس معیار دقت[10] بهترین مدل را بیابید. نمایش ماتریس پریشانی برای این مجموعه داده الزامی می‌باشد. در شکل زیر ویژگی‌های مختلف این مجموعه داده نشان داده شده است. شما می‌توانید به صورت شهودی و یا با استفاده از تکنیک‌های استخراج ویژگی بهترین ویژگی‌ها را بیابید.

توضیحات در رابطه با مجموعه داده: این مجموه داده شامل اطلاعات ۱۵۵ بیمار می‌باشد که مشکوک به بیماری هپاتیت می‌باشند. هر کدام ازاین رکوردها حداکثر دارای ۲۰ ویژگی می‌باشند (برخی از رکوردها دارای مقادیر گم شده می‌باشد که رویکرد شما در برخورد با این مقادیر بر عهده خودتان می‌باشد.). در این مجموعه داده شما باید این بیمارها را بر اساس ویژگی‌های هر کدام به دو کلاس live یا die کلاس‌بندی نمایید.

Data Dictionary

Column Position	Atrribute Name	Definition	Data Type	Example	% Null Ratios
1	Class	Class (1: DIE, 2: LIVE)	Quantitative	1, 2	0
2	Age	Age (In Years)	Quantitative	34, 20, 55	0
3	Sex	Sex (1: Male, 2: Female)	Quantitative	1, 2	0
4	Steroid	Steroid (No: 1, Yes: 2)	Quantitative	1, 2	1
5	Antivirals	Antivirals (No: 1, Yes: 2)	Quantitative	1, 2	0
6	Fatigue	Fatigue (No: 1, Yes: 2)	Quantitative	1, 2	1
7	Malaise	Malaise (No: 1, Yes: 2)	Quantitative	1, 2	1
8	Anorexia	Anorexia (No: 1, Yes: 2)	Quantitative	1, 2	1
9	Liver Big	Liver Big (No: 1, Yes: 2)	Quantitative	1, 2	10
10	Liver Firm	Liver Firm (No: 1, Yes: 2)	Quantitative	1, 2	11
11	Spleen Palpable	Spleen Palpable (No: 1, Yes: 2)	Quantitative	1, 2	5
12	Spiders	Spiders (No: 1, Yes: 2)	Quantitative	1, 2	5
13	Ascites	Ascites (No: 1, Yes: 2)	Quantitative	1, 2	5
14	Varices	Varices (No: 1, Yes: 2)	Quantitative	1, 2	5
15	Bilirubin	Bilirubin	Quantitative	0.39, 0.80, 1.20	6
16	Alk Phosphate	Alk Phosphate	Quantitative	33, 80, 120	29
17	Sgot	SGOT	Quantitative	13, 100, 200	4
18	Albumin	Albumin	Quantitative	2.1, 3.0, 3.8	16
19	Protime	Protime	Quantitative	60, 70, 80	67
20	Histology	Histology (No: 1, Yes: 2)	Quantitative	1, 2	0

[10] Accuracy

۳. مسئله خوشه‌بندی:

بر روی مجموعه داده مشخص شده زیر، حداقل ۴ الگوریتم مختلف را آزمایش نمایید و سعی کنید با استفاده از تکنیک‌های یادگرفته شده در این درس تعداد کلاسترهای بهینه را پیدا کنید. در نهایت خوشه‌بندی خود از مجموعه داده را به همان ترتیب اولیه در قالب یک فایل csv ذخیره نمایید. این فایل در روز ارائه مورد آزمایش قرار گرفته خواهد شد.

توضیحات در رابطه با مجموعه داده: این مجموعه داده شامل ۵ ویژگی کمی مربوط به عملکرد دانش‌اموزان می‌باشد. این ویژگی‌ها در شکل زیر نمایش داده شده‌اند. با توجه به این ویژگی‌ها شما می‌توانید این دانش‌اموزان را به چند خوشه که نشان دهنده‌ی سطح آنان می‌باشد تقسیم نمایید.

Data Dictionary

Column Position	Atrribute Name	Definition	Data Type	Example	% Null Ratios
1	STG	The degree of study time for goal object materials	Quantitative	0.060, 0.100, 0.080	0
2	SCG	The degree of repetition number of user for goal object materials	Quantitative	0.000, 0.100, 0.250	0
3	STR	The degree of study time of user for related objects with goal object	Quantitative	0.10, 0.15, 0.05	0
4	LPR	The exam performance of user for related objects with goal object	Quantitative	0.98, 0.10, 0.01	0
5	PEG	The exam performance of user for goal objects	Quantitative	0.66, 0.56, 0.33	0