به نام خدا

یادگیری ماشین برای بیوانفورماتیک

نیم‌سال دوم ۹۹ـ۰۰

مدرس: دکتر سلیمانی — دکتر شریفی

دانشکده مهندسی کامپیوتر

تمرین سری اول | **مقدمات** | موعد تحویل: ۲۸ اسفند

مسئله‌ی ۱.(۱۰ نمره)

1. **(۶نمره)** متغیر تصـادفی X دارای توزیع $f_X(x) = \begin{cases} cx & 0 \le x \le 1 \\ c(2-x) & 1 \le x \le 2 \\ 0 & o.w \end{cases}$ اسـت. مقادیر زیر را برای این توزیع محاسبه کنید.
   - ضریب c
   - CDF
   - Expected value
2. **(۴نمره)** متغیرهای تصادفی مستقل $X_1, X_2, \cdots, X_n$ با توزیع هندسی

$$P(X = k) = \theta(1-\theta)^{k-1} \quad k \in \{1, 2, 3, \cdots\}$$

را در نظر بگیرید. MLE را برای پارامتر θ به دست آورید.

مسئله‌ی ۲.(۱۰ نمره)

1. **(۶نمره)** اگر A ماتریس مربعی، b یک متغیر ، a و x بردارهای ستونی باشند، عبارت‌های زیر را ثابت نمایید:
   - $\frac{da^T x}{dx} = \frac{dx^T a}{dx} = a^T$
   - $\frac{d(x^T a)^2}{dx} = 2x^T a a^T$
   - $\frac{dx^T Ax}{dx} = x^T(A + A^T)$
2. **(۴ نمره)** مقدار ویژه و بردار ویژه را برای ماتریس HH^T را بدست آورید:

$$H = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

مسئله‌ی ۳.(۱۰ نمره)

فرض کنید n داده آموزش به صورت $D = (x^1, y^1), \cdots, (x^n, y^n)$ در اختیار داریم که هر کدام از x ها، d بعدی می باشد. می خواهیم از رگرسیون خطی با تابع هزینه SSE استفاده کنیم که به فرم زیر است:

$$J(\omega) = \sum_{i=1}^{n} (y^{(i)} - \omega^T x^{(i)})^2$$

1. **(۲ نمره)** رابطه بهینه برای ω را به دست آورید.
2. **(۳ نمره)** مشکلات استفاده مستقیم از رابطه قسمت قبل را بیان کنید و برای آن ها راه حلی ارائه دهید.
3. **(۲نمره)** اگر به تابع هزینه جمله منظم‌ساز $\|\omega\|^2$ را بیفزاییم، فرم بسته پاسخ بهینه ω را به دست آورید.
4. **(۳ نمره)** رگرسیون خطی وزن دار، تعمیمی از رگرسیون خطی است که در آن به هر یک از داده ها، وزنی اختصاص داده می‌شود:

$$J(\omega) = \sum_{i=1}^{n} f_i (y^{(i)} - \omega^T x^{(i)})^2$$

فرم بهینه ω را برای این تابع هزینه به دست آورید.

مسئله‌ی ۴.(۱۵ نمره)

اطلاعات تشخیص بیماری براساس تعدادی از ژن‌هایی که فکر می‌کنیم بر روی بیماری اثر گذار هستند در جدول زیر آمده است. از آن‌جایی که اطلاعات این جدول به صورت عددی آمده است، ابتدا با استفاده از threshold به ۲ حالت زیاد و کم تبدیل کنید و به سوالات زیر پاسخ دهید. Threshold تعیین شده برای ژن۱ (۱۰۰) ، برای ژن۲ (۵۰) ، برای ژن۴ (۵) ، برای ژن۵ (۲۵) است.

۱ ـ جدول داده‌های آموزش

نام شخص	ژن۱	ژن۲	ژن۳	ژن۴	ژن۵	بیمار بودن؟!
X1	۱۰	۳۰	زیاد	۱	۴۰	بله
X2	۱۰۱	۳۱	متوسط	۰	۱۴	خیر
X3	۲۰	۷۰	زیاد	۲	۳۳	بله
X4	۲۰۰	۳۵	کم	۳	۱۰	خیر
X5	۳۷	۹۰	متوسط	۲۲	۱۱	بله
X6	۲۵	۲۵	متوسط	۱۹	۵۰	خیر
X7	۴۰	۷۵	زیاد	۳۰	۵۵	خیر
X8	۵۰	۴۵	کم	۴	۱۲	بله
X9	۱۲۰	۸۰	متوسط	۲۳	۸	خیر

۲ ـ جدول داده‌ها تست

نام شخص	ژن۱	ژن۲	ژن۳	ژن۴	ژن۵	بیمار بودن؟!
Y1	۴۲	۲۰	زیاد	۰	۳۷	بله

Y2	۱۳۹	۸۵	کم	۱۵	۵۰	خیر
Y3	۵۹	۴۱	زیاد	۴	۵۰	بله
Y4	۲۲	۳۷	متوسط	۱	۸۰	خیر
Y5	۷۶	۶۵	متوسط	۲	۶۰	بله
Y6	۳۰	۸۳	زیاد	۱	۱۷	خیر

1. **(۵ نمره)** درخت تصمیم را به صورتی دستی بر روی داده‌های یادگیری آموزش دهید و دقت دسته‌بند را بر روی داده‌های تست بررسی نمایید.

* برای سوالات زیر درصورتی‌که در درخت تصمیم تغییری ایجاد شود، آن را مجددا رسم نموده و دقت دسته‌بند را برای داده‌های تست حساب نمایید. چانچه در درخت تصمیم تغییری ایجاد نشود دلیل آن را به طور کامل توضیح دهید.

2. **(۲نمره)** اگر ژن شماره ۲ وجود نداشت تغییری در درخت آموزش ایجاد می‌شد؟ در صورتی‌که ژن شماره ۳ نباشد آیا باز هم تغییری در درخت تصمیم ایجاد نمی‌شود؟
3. **(۴ نمره)** اگر فرد X5 به عنوان فرد سالم در نظر گرفته می‌شد، تغییری در درخت تصمیم رخ می‌داد؟
4. **(۴نمره)** اگر فرد جدیدی به نام X10 به داده‌های آموزش اضافه می‌شد و علاوه بر اطلاعات زیر، شامل چند ژن دیگر با مقادیر تصادفی‌کم یا زیاد بود، تغییری در درخت تصمیم ایجاد می‌کرد؟

X10	۱۷	۸۰	کم	۲۷	۲۰	خیر

مسئله‌ی ۵.(۳۵ نمره +۱۰ نمره)

کتابخانه‌های numpy و pandas کتابخانه‌هایی بسیار سریع و بهینه‌ای هستند که برای پردازش داده‌های بزرگ از آن‌ها استفاده می‌شود به همین خاطر در این سوال قصد داریم برای‌کار با داده، از این دو کتابخانه استفاده‌کنیم. برای پیاده‌سازی این سوال تنها استفاده از این دو کتابخانه مجاز است.

داده‌های یک مجموعه بیمارستانی‌که شامل اطلاعات بیماران دیابتی و غیر دیابتی بوده در اختیار ما قرار گرفته است. قصد داریم با استفاده از درخت تصمیم و Perceptron، بر روی داده بدست آمده، دیابتی بودن و یا دیابتی نبودن هر فرد را تشخیص دهیم. این داده دارای ۹ ستون اطلاعاتی بوده‌که ستون ۱ تا ۸ آن اطلاعات هر فرد و ستون آخر یعنی Outcome بیانگر بیمار بودن یا نبودن آن‌هاست.

۵/۱. (۱۷نمره + ۷ نمره) درخت تصمیم:

1. **(۲نمره)** ابتدا داده‌های آموزش و تست را خوانده و boxplot هر ویژگی را نسبت ستون آخر (Outcome) که برچسب بیمار بودن یا نبودن را نشان می‌دهد، ترسیم نمایید (برای این‌کار از کتابخانه matplotlib و seaborn استفاده نمایید).

2. **(۹نمره)** سپس با توجه به این که داده‌های موجود در ستون‌های ۱ تا ۸ عددی هستند ابتدا میانگین هر ستون را محاسبه کنید و مقادیر کمتر از میانگین را برچسب low و برای مابقی آنها از برچسب high استفاده کنید . پس از تغییر و اصلاح آن، درخت تصمیم پیاده‌سازی کرده و بر روی داده آموزش دهید. برای پیاده‌سازی درخت تصمیم و انتخاب ویژگی‌ها از معیار Information Gain استفاده کنید. از آن‌جایی که عمق درخت تصمیم یکی از هایپرپارامترهای این دسته‌بند است، آن را به صورت پارامتری نگه‌داشته تا درگام‌های بعدی بتوانید مقادیر مختلفی را روی آن امتحان نمایید.
3. **(۳نمره)** درخت تصمیم خود را روی عمق یک تا عمق ۸ (که شامل تمام ویژگی‌هاست) آموزش دهید. دقت دسته‌بند را به ازای تمامی عمق‌ها محاسبه کنید. نمودار دقت را بر روی داده‌ها آموزش و تست بر حسب محدودیت عمق درخت رسم نموده و افت و خیزهای آن را توصیف نمایید.
4. **(۳نمره)** با استفاده از 5fold-cross validation مناسب‌ترین عمق را برای درخت انتخاب کنید. سپس درخت تصمیم را با اعمال محدودیت عمق بر روی تمامی داده‌های آموزش تعلیم دهید. معیارهایf-score، sensitivity و specificity را برای داده تست گزارش کنید. ارزش این معیارها در مقابل معیار دقت چیست و چه زمانی هر کدام از آن‌ها اهمیت بیشتری پیدا می‌کنند.
5. **(+۵نمره)** درخت ساخته شده را به بهترین صورت ممکن هرس نمایید و نتیجه آن را با حالت قبل مقایسه کنید. چه نتیجه‌ای می‌گیرید؟
6. **(+۲نمره)** برای بخش قبلی، دقت حالت هرس شده و هرس نشده را از طریق paired t-test مقایسه و نتایج تست را تفسیر نمایید.

۵/۲.(۱۸ نمره + ۳ نمره) Perceptron:

1. **(۹نمره)** دسته‌بند Perceptron به صورت ساده پیاده‌سازی کرده و سپس این دسته‌بند را بر روی داده بکار بگیرید .
   - ابتدا مجددا داده خام را load کرده و ردیف‌هایی که مقادیر هر یک از ستون‌های "BMI" ، "Glucose" ، "BloodPressure" برای آن‌ها صفر است را از داده حذف نمایید. برچسب Outcome را به صورت ۱ـ و ۱ در نظر بگیرید. داده‌ها موجود در هر ستون را پیش از استفاده به کمک max و min هر ویژگی نرمال نمایید.
   - داده‌ها آموزش را به دو بخش train و validation تقسیم کنید(۸۵ درصد داده آموزش به بخش train و ۱۵ درصد آن به بخش validation اختصاص داده شود) و در انتها نمودار تغییرات دقت را برای داده validation رسم کنید.
   - بهترین دسته‌بند را بر اساس دقت انتخاب نمایید و معیارهایf-score، sensitivity و specificity را بر روی داده‌های تست گزارش کنید.
2. **(۹ نمره + ۳نمره)** دسته‌بند Perceptron به صورت full batch و mini batch پیاده‌سازی کرده و نمودار تغییرات دقت را برای داده validation رسم و معیارهایf-score، sensitivity و specificity را بر روی داده‌های تست گزارش کنید.

مسئله‌ی ۶.(۲۰ نمره)

در این سوال میخواهیم از رگرسیون خطی برای پیش‌بینی هزینه پزشکی افراد بر اساس ویژگی‌های شخصی آن‌ها استفاده کنیم. توجه کنید که این سوال را باید از پایه و به کمک numpy و pandas پیاده‌سازی کنید و استفاده از کتابخانه های آماده یادگیری ماشین مجاز نیست. داده های مربوط به این سوال در پوشه regression قرار دارد. هر داده دارای ۶ ویژگی ورودی (x) و یک خروجی هزینه پزشکی (y) است. همچنین ویژگی های جنسیت، منطقه و سیگاری بودن از نوع categorical هستند که باید اصلاح شوند. با توجه به این که جنسیت و سیگاری بودن، تنها دو مقدار مشخص میگیرند، میتوانید مقادیر آن ها را با ۰ یا ۱ جایگزین کنید (integer encoding) اما برای منطقه، باید از one hot encoding استفاده کنید.

1. **(۶ نمره)** رگرسیون خطی تعمیم یافته را بدون جمله منظم ساز پیاده‌سازی کرده و نتایج را روی داده‌های تست گزارش کنید. با توجه به این که نشان داده شده است که هزینه‌های پزشکی با سن رابطه خطی ندارد و با افزایش سن، نرخ افزایش هزینه ها بیشتر میشود ، تابع *basis* برای ویژگی سن را[1] به صورت $\phi(x) = x^2$ و برای بقیه ویژگی‌ها به صورت همانی $\phi(x) = x$ در نظر بگیرید. در این قسمت از فرمول بسته رگرسیون خطی تعمیم یافته برای محاسبه w استفاده کنید و مقدار نهایی را در گزارش یادداشت کنید.
2. **(۶ نمره)** روش *SGD* را پیاده‌سازی کنید و مقدار w و خطا را گزارش کنید.
3. **(۸ نمره)** به تابع هزینه ی *SSE* جمله ی منظم ساز L_2 با ضریب λ بیفزایید. با استفاده از *5-fold cross validation*، بهترین مقدار پارامتر λ را از بین $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}\}$ بیابید و سپس نتیجه را روی داده‌های تست بدست آورید. نمودار خطا را برحسب لگاریتم λ رسم کنید. در این قسمت مجددا از فرمول بسته ی w استفاده کنید. نتایج نهایی و مقدار خطا را برای داده های تست و آموزش در گزارش بنویسید.

[1] Polder JJ, Bonneux L, Meerding WJ, van der Maas PJ. Age-specific increases in health care costs. Eur J Public Health. 2002 Mar;12(1):57-62. doi: 10.1093/eurpub/12.1.57. PMID: 11968522.