به نام خدا

دانشکده‌ی علوم و مهندسی کامپیوتر

یادگیری ماشین

تکلیف اول – پاییز ۱۳۹۴

قبل از اینکه شروع به حل تمرین‌ها کنید، حتما یک بار فایل Policies.pdf (موجود در وبگاه) که حاوی نکات مهم در تحویل تکالیف هست را مطالعه فرمایید.

حل سوالاتی که امتیازی هستند، اجباری نبوده و نمره‌ی اضافی خواهد داشت.

طراح هریک از سوالات در زیر مشخص شده است .در صورتی که سوالی دارید می توانید سوالتان را در گروه مطرح کنید یا از طراح سوال بپرسید.

۱ الی ۴: آقای زندی

۵ الی ۸: آقای کاهانی

۱- فرض کنیم که شما یک ورزشکار هنرهای رزمی هستید و متغیر تصادفی x نشان‌دهنده‌ی باخت (۰) یا برد (۱) شما در مقابل یک حریف است. قرار است که شما با یکی از سه ورزشکار قدرمتند جهان مبارزه دهید. اما نمی‌دانید که دقیقا کدام! هر یک از حریف‌های شما نقاط ضعف خاص خود را دارند و در نتیجه توزیع احتمال برد یا باخت شما در برابر هر یک از آنها متفاوت است. به علاوه احتمال اینکه با کدام حریف مواجه شوید نیز تفاوت دارد.

	احتمال مواجه شدن	احتمال برد شما
قهرمان ۱	0.3	0.1
قهرمان ۲	0.5	0.4
قهرمان ۳	0.2	0.8

الف- فرض کنید که متغیر تصادفی $\theta = \{1,2,3\}$ نشان‌گر حریف شما باشد. $p(x|\theta)$ ، $p(x)$ و $p(\theta|x)$ را محاسبه کنید و توضیح دهید که هر کدام از این توابع احتمال نشان‌گر چیست.

ب- اگر که شما مسابقه را برده باشید، استنتاج کنید که احتمالا کدام قهرمان حریف شما بوده است؟

ج- اگر به ازای برد ۱ میلیون تومان جایزه ببرید ولی به ازای باخت ۲ میلیون ضرر کنید، منطقی است که در این مسابقه شرکت کنید یا خیر؟

۲- تابع چگالی احتمال توزیع پواسون با پارامتر $\lambda>0$ به صورت زیر است:

$$f(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0,1,2,...$$

الف- با استفاده از تکنیک maximum likelihood، مقدار بهینه‌ی پارامتر λ را حساب نمایید. فرض کنید، اعضای مجموعه‌ای که می‌خواهید برای آن پارامتر بهینه را حساب کنید، مستقل هستند و همچنین همگی از یک توزیع پواسون استخراج شده‌اند. تعداد آنها را نیز n فرض کنید.

ب- به صورت ریاضی اثبات نمایید که پارامتر تخمین زده شده، دارای بایاس نیست.

ج- واریانس پارامتر تخمین زده شده، $\frac{\lambda}{n}$ است. این رابطه را اثبات نمایید.

د- آیا این تخمین، یک تخمین سازگار است؟

ه- اگر توزیع λ به صورت زیر باشد، مقدار بهینه‌ی پارامتر λ را به دست آورید.

$$p(\lambda)=\frac{\lambda}{100},\ 5\le\lambda\le 15$$

۳- (امتیازی) تابع چگالی احتمال توزیع نمایی به صورت زیر را در نظر بگیرید: ($\lambda>0$)

$$f(x\mid\lambda)=\begin{cases}\lambda e^{-\lambda x} & x\ge 0\\ 0 & x<0\end{cases}$$

اگر از روش Maximum Likelihood برای تخمین پارامتر λ استفاده کنیم، مقدار بایاس آن را حساب نمایید. روابط زیر در این اثبات می‌توانند مفید باشند: (فرض i.i.d بودن نمونه‌ها را نیز در نظر بگیرید)

- $\bar{x}_n=\sum_{i=1}^{n}x_i,\ x_i\sim Exp(\lambda)\Rightarrow\bar{x}_n\sim Erlang(n,\lambda)$
- اگر $f(x)$ تابع چگالی احتمال متغیر تصادفی x باشد: $E\left[g(x)\right]=\int_{-\infty}^{\infty}g(x)\,f(x)\,dx$
- $\int_0^{\infty}x^n e^{-ax}dx=\frac{n!}{a^{n+1}}$

۴- بعد VC را برای یک بیضی در فضای دو بعدی حساب نمایید. فرض کنید که مکان، اندازه و زاویه چرخش بیضی همگی قابل یادگیری هستند. با شکل نشان دهید که n داده را چطور در صفحه در نظر می‌گیرید.

۵- دو کلاس C_1 و C_2 با تک ویژگی x را داریم. هزینه ی انتخاب اشتباه کلاس ۱ برابر 0.5 و کلاس ۲ برابر m است. اگر توزیع داده‌های کلاس‌ها، گاوسی با پارامترهای $N_1(2, 1)$ و $N_2(4, 1)$ و احتمال پیشین $P(C_1) = 0.6$ و $P(C_2) = 0.4$ باشد، مقدار m را طوری تعیین کنید که مرز جداساز در نقطه ی $x = L$ قرار گیرد.

۶- الف) معادله‌ی صفحه ی جداساز دسته‌بند بیزی برای دسته بندی دو کلاس C_1 و C_2 با ۳ ویژگی x, y, z را بدست آورید. (توزیع هر دو کلاس را گاوسی در نظر بگیرید.)

ب- مجموعه داده ای از بیماران سرطانی جمع آوری شده در سال های ۱۹۵۸ تا ۱۹۷۰در اختیار داریم، این مجموعه دارای ۳۰۶ نمونه بوده که هر نمونه ۴ ویژگی دارد. می خواهیم ویژگی آخر را که وضعیت زنده ماندن یا مردن بیمار را ۵ سال بعد از عمل جراحی نشان می دهد، با استفاده از دسته‌بند بیزی پیش‌بینی نماییم. این مجموعه در فایل haberman.data قرار دارد.

با استفاده از جداساز بدست آمده در قسمت قبل (با فرض گاوسی بودن توزیع کلاس‌ها)، دسته بندی را انجام داده و دقت را محاسبه نمایید.

توضیح: برای دسته بندی از متد اعتبار سنجی متقابل k-fold با $k = 4$ استفاده نمایید.

ج) این بار پیچیدگی مدل را کاهش داده و ماتریس کواریانس کلاس‌ها را مشترک در نظر بگیرید. با استفاده از معادله ی جداساز بدست آمده در قسمت الف، نمونه های دو کلاس و جداساز مربوطه در هر بار دسته‌بندی را رسم نمایید.

توضیح: در این بخش نیز از روش k-fold با $k = 4$ برای مشخص نمودن داده های آموزشی استفاده نمایید.

د) این‌بار از تمام نمونه ها برای آموزش استفاده کرده و جداساز بدست آمده را رسم کنید. همچنین میانگین مدل های بدست آمده در بخش قبل را نیز رسم کرده و نتایج را مقایسه نمایید.

توضیح: در این بخش نیز لازم است ماتریس کواریانس‌ها مشترک در نظر گرفته شوند.

۷- مجموعه داده‌ای از قیمت خانه و سه ویژگی تاثیرگذار در تعیین قیمت آن داده شده است. این مجموعه داده شامل ۱۵۰ نمونه و ۴ ویژگی است که در فایل house.txt قرار داده شده است.

الف- اگر بخواهیم قیمت خانه را تنها با استفاده از یکی از سه ویژگی و با رگرسیون خطی پیش‌بینی نماییم، چگونه می‌توان بهترین ویژگی را قبل از محاسبه‌ی رگرسیون تعیین کرد؟ ویژگی‌ها را بر این اساس رتبه‌بندی نمایید.

ب) با استفاده از مجموعه داده‌ی خانه، این‌بار با استفاده از هر سه ویژگی یک مدل خطی برای تخمین قیمت خانه بدست آورید. سپس مدل را رسم نموده و میزان خطای آن را محاسبه نمایید.

۸- اگر از تابع $y = x^2$ برای x در بازه ی $[-4, 4]$، نمونه برداری با فواصل یکسان انجام دهیم، آیا رابطه ای خطی بین x و y وجود دارد؟ آیا می توان با استفاده از رگرسیون مدلی خطی برای تقریب این رابطه تخمین زد؟ اگر بله، روش پیشنهادی را ذکر کرده و نتیجه را رسم نمایید.

موفق باشید