به نام خدا

یادگیری ماشین

تکلیف نخست– پاییز ۱۳۹۶ مهلت ارسال: ۲۰ آبان ۱۳۹۶

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 دستیار درس بپرسید.

۱- با توجه به شکل زیر به سوالات پاسخ دهید:

شکل ۱

الف) همان‌طور که می‌دانید تعریف‌های گوناگونی برای یادگیری ماشین ارائه شده است. با توجه به گستردگی تعاریف، تعریفی از یادگیری ماشین ارائه کنید که بتواند به بهترین صورت آنالیز داده‌های شکل ۱ را به عنوان کاربردی از یادگیری ماشین توصیف کند.

ب) برای شکل ۱ مساله ای طراحی کنید و با استفاده از یادگیری ماشین برای آن راه حلی پیشنهاد نمایید.

ج) مساله تشخیص خط را درنظر بگیرید، یادگیری ماشین چه کمکی در بدست آوردن راه حل می‌تواند انجام دهد؟

د)آیا می‌توانیم ادعا کنیم در حل مساله تشخیص خط به یادگیری ماشین نیازمندیم؟

۲- عموما بین پیچیدگی فضای فرضیه، مقدار خطای generalization و تعداد داده‌های آموزشی رابطه‌ای متناظر وجود دارد بیان کنید تغییرات هریک چه اثری بر دیگری دارد.

۳- روش یادگیری Naive Bayes Classifier بر پایه این فرض ساده (Naive) عمل می‌کند که:

ویژگی‌های مختلف مستقل هستند.

Naïve Bayes classifier با استفاده از رابطه ای که بین tokenها (برای نمونه کلمات) وجود دارد به منظور spam filtering استفاده می‌شود، به این شکل که بتواند با استفاده از تکنیک دسته‌بندی bayes با محاسبه احتمال spam بودن یا نبودن یک email اقدام به جداسازی آن‌ها نماید. حال فرض کنید پیام های مشکوک آلوده با کلمه "replica" باشد. بیشتر افرادی که ایمیل دریافت می‌کنند می‌دانند وجود چنین کلمه‌ای نشان دهنده احتمال بالای spam بودن آن است. برای نمونه می‌تواند یک پیشنهاد جهت فروش کپی جعلی یک ساعت با برند شناخته شده باشد. یک spam detection software عموما از وجود بسیاری از جزئیات اطلاعی ندارد از این رو اقدام به محاسبه احتمال می‌نماید.

Pr(S): احتمال spam بودن هر پیام.

Pr(W|S): احتمال اینکه کلمه "replica" در پیام های spam دیده شود.

Pr(H): احتمال اینکه هر پیامی که دیده شده است spam نباشد.

Pr(W|H): احتمال اینکه کلمه "replica" در پیام های سالم دیده شود.

الف) فرض کنید Pr(S|W) احتمال spam بودن پیام به در میان پیام های حاوی کلمه "replica" باشد. با در نظر گرفتن توابع احتمالاتی مورد نیاز به صورت پارامتری احتمال Pr(S|W) را محاسبه نمایید.

ب) حداقل ۴ دلیل ذکر کنید که چرا در نظر گرفتن فرض idd[۱] باعث مختل شدن عملکرد spam filter می‌شود. و در هرکدام نحوه ایجاد اختلال spamها را توضیح دهید.

ج) براساس پیاده‌سازی که جهت طراحی spam filter انجام می‌دهید، Spam filter شما ممکن است به spamها حساسیت‌های مختلفی نشان دهد یا حتی اصلا حساس نشود. کسانی که spamها را تولید می‌کنند سعی می‌کنند به طریقی این حساسیت را کاهش دهند تا بتوانند از این موانع عبور کنند. به عنوان یک فروشنده جنس قلابی! سعی کنید ایده‌هایی جهت عبور از این spam filterها ارائه دهید.

[۱] Independent and identically distributed random variables

۴- فرض کنید در یک غار تاریک با ۳ خروجی غیر قابل تشخیص و پنهان در دیوارها گیر افتاده اید. یکی از درها به مسیری ۳ ساعته برای خروج به سمت بیرون ختم می‌شود. دو در دیگر نیز دارای مسیری است که شما را در طی ۱ و ۲ ساعت دوباره به غار برمی‌گرداند. شما راهی برای تشخیص تمایز درها ندارید. تخمین بزنید چقدر طول می‌کشد به بیرون برسید.

۵- فرض کنید داده های $X_1,\ldots,X_n$ داده هایی idd از یک توزیع یکنواخت حول دایره ای به شعاع θ در فضای R^2 باشند به طوریکه $X_i \in R^2$ و

$$P(x\,|\,\theta)=\begin{cases}\dfrac{1}{\pi\theta^2} & \text{if } |x|\le\theta\\ 0 & otherwise\end{cases}$$

با استفاده از MLE، θ را تخمین بزنید.

۶- (امتیازی)در این سوال با برنامه نویسی در محیط Python آشنا خواهید شد. برای شروع کدهای زیر را در محیط برنامه کپی نمایید سپس گام به گام مراحل خواسته شده در زیر را طی کنید. داده‌هایی که مورد استفاده قرار می‌گیرد داده‌های "Fisher iris" می‌باشد این داده‌ها شامل ۴ اندازه حقیقی به عنوان ویژگی‌های سه نوع گل زنبق(iris) می‌باشد.

```
import numpy as np
import matplotlib.pyplot as plt

iris = np.genfromtxt("data/iris.txt",delimiter=None)  # load the text file
Y = iris[:,-1]              # target value is the last column
X = iris[:,0:-1]            # features are the other columns
```

الف) با استفاده از برنامه نویسی تعداد ویژگی ها و تعداد داده ها را بدست آورید.

ب) برای هر ویژگی هیستوگرام مقادیر داده ها را رسم کنید.

ج) برای هر ویژگی میانگین آن‌ها را محاسبه کنید

د) برای داده‌های هر ویژگی مقادیر واریانس و انحراف معیار را بدست آورید.

ه) داده ها را با محاسبه تفاضل میانگین از داده ها بخش بر انحراف معیار نرمال نمایید.

۷- روش Bayes یکی از راه‌های بدست آوردن مرز تصمیم در دسته‌بندی‌هاست[۲]. در این حالت ما مرز تصمیم بهینه را برای دو کلاس C_1 و C_2 با استفاده از نسبت likelihood به دست می‌آوریم.

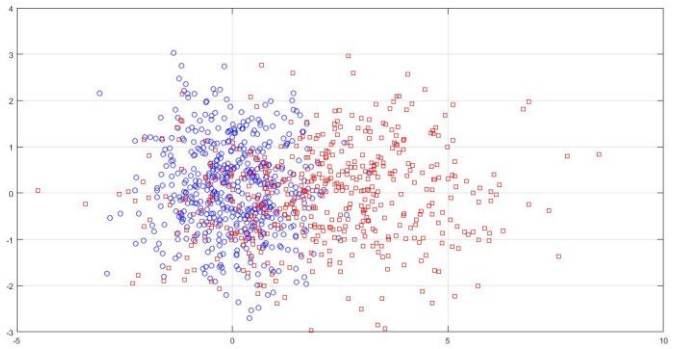
داده های Class1 و Class2 مربوط به شکل زیر را درنظر بگیرید هریک شامل ۵۰۰ داده دوبعدی است. قصد داریم این داده‌ها را به وسیله یک مرز از هم جدا سازیم:

فرض کنیم می‌دانیم، توزیع دادها درکلاس C_1 بصورت گوسی با بردار میانگین [0,0] و واریانس ۱ و برای کلاس C_2 با بردار میانگین [0,2] و واریانس ۴ می‌باشد. و threshold ما ۱ باشد.

الف) ابتدا داده‌ها را در محیط متلب وارد کنید و شکل هرکدام را ترسیم کنید.

ب) با استفاده از تکنیک تخمین با استفاده از بیشینه‌ی درست‌نمایی سعی کنید مرزی برای جداسازی این دو کلاس را به صورت دستی و با کد متلب بیابید.

ج) به نظرتان مرز بدست آمده بهینه است.



شکل ۲-پراکندگی داده ها با هر دو کلاس

[۲] Bayesian Decision Boundary