# دانشگاه تهران

## دانشکده علوم مهندسی

## الگوریتم‌ها و محاسبات

**یادگیری ماشین – دکتر سایه میرزایی**

**تمرین دوم**

بهار ۰۰

## مدل‌های مولد[1]

### سوال اول:

**۱**- طبقه‌بندی با استفاده از مدل‌های مولد را به طور کامل شرح دهید، سپس روابط استفاده شده برای این طبقه‌بند را به صورت **شهودی** تفسیر کنید.

**۲**- مزایا و معایب استفاده از این طبقه‌بندی را شرح دهید. برای حل یک مسئله یادگیری ماشین، چه‌زمانی استفاده از طبقه‌بند مولد را پیشنهاد می‌کنید؟

**۳**- آیا می‌توان از این طبقه‌بند برای وقتی که ازهمه ویژگی‌های دادگان تست به صورت کامل اطلاع نداریم استفاده کرد؟ چرا؟

**۴**- مدل $GDA$[2] و روابط آن را تشریح کنید و آن را با رگرسیون لاجستیک مقایسه کنید.

**۵**- مدل $LDA$[3] و $QDA$[4] را بررسی کنید و روابط آن‌ها را شرح دهید. این دو مدل چه تفاوتی با هم دارند؟

### سوال دوم:

در این تمرین می‌خواهیم داده‌های سرطان سینه را با استفاده از الگوریتم $Naïve\ Bayes$ آموزش دهیم و بیمارهای بدخیم و خوش‌خیم را تشخیص دهیم.

**۱**- فرض کنید تمام ویژگی‌ها از توزیع نرمال با واریانس و میانگین متفاوت برای بیماری بدخیم و خوش‌خیم پیروی کنند، حال با استفاده از روش $Naïve\ Bayes$ داده‌ها را آموزش دهید، در این قسمت از ۲۰ درصد داده‌های برای تست کردن مدل خود استفاده کنید و سپس دقت مدل خود را گزارش کنید.

**۲**- با توجه به قسمت قبل اگر جواب آزمایش یک فرد خوش‌خیم بودن را نشان دهد با چه احتمالی بیماری خوش‌خیم است؟

### سوال سوم:

در این سوال با داده‌های هزینه خانه‌ها در بوستون کار خواهیم کرد.

**۱**- برای طبقه‌بندی به کمک طبقه‌بند بیزین می‌توان از دو تخمین زن $MLE$[5] و $MAP$[6] استفاده کرد. این دو تخمین زن را با ذکر روابط و یک مثال شرح دهید و سپس با هم مقایسه کنید و مزایا و معایب هر کدام را بنویسید.

---

[1] Generative model
[2] Gaussian Discriminant Analysis
[3] Linear Discriminative Analysis
[4] Quadratic Discriminant Analysis
[5] Maximum Likelihood Estimation
[6] Maximum A Posteriori Estimation

**۲**- میانگین را برای داده‌های هدف پیدا کنید، این داده‌ها نشان‌دهنده قیمت خانه‌های بوستون هستند، داده‌های بیشتر از میانگین را به‌عنوان گران قیمت (با عدد ۱) و داده‌های کمتر از میانگین را ارزان قیمت (با عدد صفر) برچسب‌گذاری کنید و جایگزین داده‌های هدف کنید.

**۳**- ۲۰ درصد داده‌ها را به‌عنوان داده‌های تست در نظر بگیرید، پیش‌پردازش‌های مورد نیاز را روی داده‌ها انجام دهید و سپس با استفاده از تخمین زن $MLE$ کلاس داده‌های تست را تخمین بزنید و دقت مدل را به همراه **ماتریس درهم‌ریختگی**[۷] گزارش کنید.

**۴**- ۶۰ درصد داده‌ها را به‌عنوان داده‌های تست در نظر بگیرید و سپس قسمت ۳ را تکرار کنید و نتایج را مقایسه کنید.

**۵**- راهکاری برای حل مشکل قسمت ۴ پیشنهاد کنید. (پیاده‌سازی این قسمت اختیاری می‌باشد و امتیاز مثبت دارد)

## نکات

- **تمرین‌ها را در سامانه ایلرن تحویل بدهید.**
- لطفا گزارش خود را به زبان فارسی تهیه کنید و تمامی نکات، فرض‌ها و فرمول‌ها در آن ذکر شوند. گزارش در روند تصحیح تمرین‌ها از اهمیت ویژه‌ای برخوردار است.
- کپی‌کردن کدهای آماده موجود در اینترنت و یا استفاده از کدهای همکلاسی‌ها تقلب محسوب می‌شود.
- استفاده از کتابخانه‌های آماده پایتون به جز $Numpy$، $Pandas$ و $Matplotlib$ غیرمجاز است، تنها برای بارگذاری داده‌ها mat. می‌توانید از کتابخانه‌های دیگر استفاده کنید.
- در صورت مشاهده تقلب نمرات تمامی افراد شرکت‌کننده در آن صفر لحاظ می‌شود.
- پس از به اتمام رسیدن مهلت تحویل تمرین، تاخیر تا یک هفته با کسر ۳۰ درصد نمره لحاظ خواهد شد.
- در صورت وجود هرگونه ابهام یا مشکل می‌توانید از طریق گروه کلاسی یا ایمیل mo.bakhtyari@ut.ac.ir با دستیار آموزشی در تماس باشید.

[7] Confusion Matrix