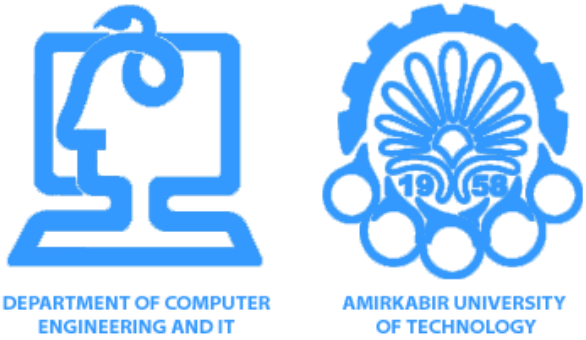


دانشگاه صنعتی امیرکبیر
دانشکده مهندسی کامپیوتر

تمرین سوم 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 ۲۷ آذر** می‌باشد **و به هیچ وجه تمدید نمی‌شود**.

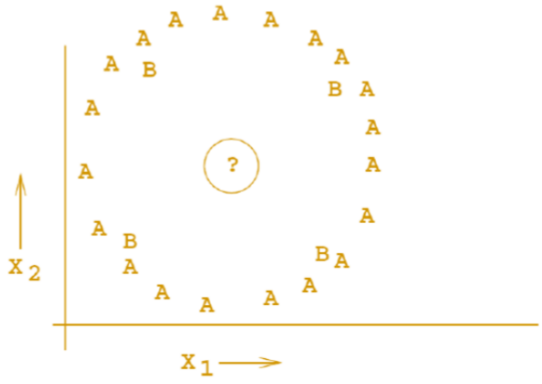
تمرینات تشریحی: (۳۰٪)

۱- با استفاده از مراجع [۱] و [۲] و سایر مراجع در صورت نیاز، دسته بند بیز ساده و رگرسیون لاجیستیک[1] را با یکدیگر مقایسه کنید. (حداکثر در یک صفحه)

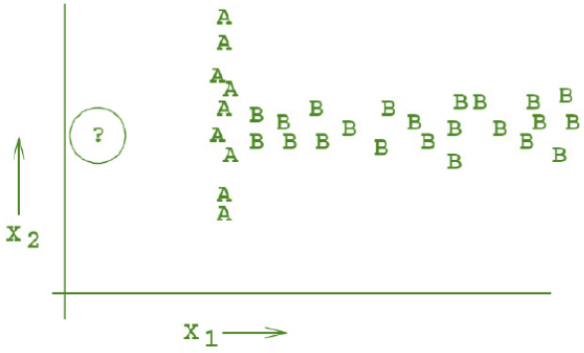
۲- توضیح دهید که عملیات هموارسازی[2] در دسته بند بیز ساده به چه منظور انجام می‌شود و در مورد روش‌های مختلف آن به صورت مختصر توضیح دهید. (با ذکر مرجع)

۳- فرض کنید می‌خواهیم دسته بند ساده گاوسی را برای $y = A \wedge B$ که A و B مستقل هستند آموزش دهیم. می دانیم که $p(A) = 0.3$ و $p(B) = 0.6$. گراف و جدول رخداد مربوط به این دسته‌بند را رسم کنید.

۴- فرض کنید برای هر کدام از اشکال ۱و ۲ یک دسته‌بند ساده گاوسی آموزش داده‌ایم. توضیح دهید که برچسب داده تست که با علامت سوال مشخص شده است چه خواهد بود.



شکل ۱

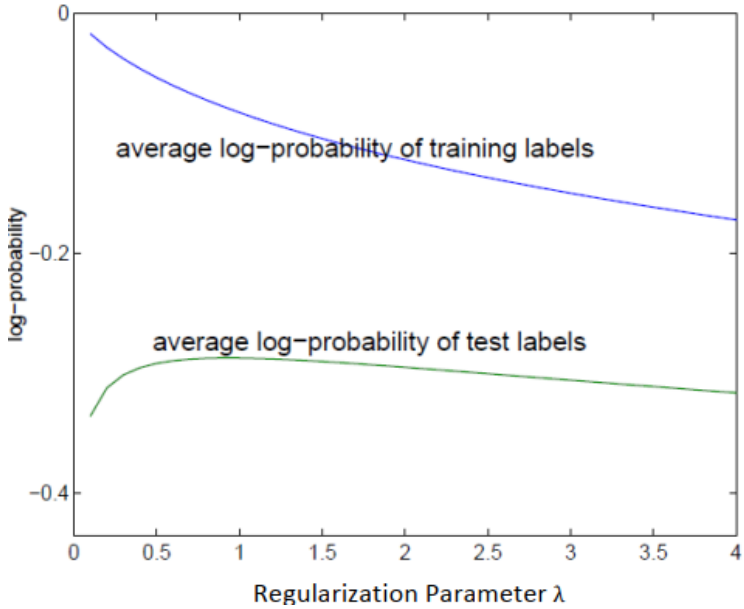


شکل ۲

۵- تابع هزینه مورد استفاده در رگرسیون لاجیستیک در حالت منظم شده[3] در نظر بگیرید. نقش پارامتر منظم سازی را در این تابع هزینه بررسی کنید. و با توجه به تصویر ۳ توضیح دهید که چرا با افزایش پارامتر منظم سازی میانگین لگاریتم احتمال برچسب داده‌ها کاهش می‌یابد.

---

[1] Logistic Regression
[2] Smoothing
[3] Regularized

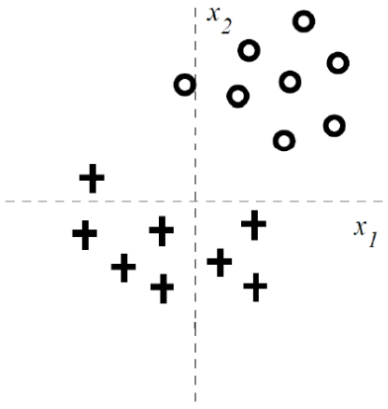


شکل ۳

۶- داده‌های موجود در شکل ۴ مدل رگرسیون لاجیستیک را با تابع هزینه منظم‌شده‌ایکه به صورت زیر است در نظر بگیرید.

$$j(\theta) = -\left[\frac{1}{m}\sum_{i=1}^{m} y^{(i)}\log h_\theta\left(x^{(i)}\right) + \left(1-y^{(i)}\right)\log\left(1-h_\theta\left(x^{(i)}\right)\right)\right] + \frac{\lambda}{2m}\theta_j^2$$

به عبارت دیگر در این رابطه فقط یکی از پارامترها تنظیم می‌شود. توضیح دهید که با در نظر گرفتن هرکدام از پارامترهای θ_j به ازای مقدار بزرگ λ خطای دسته‌بندی چگونه تغییر می‌کند.



شکل ۴

در انجام تمرینات زیر به نکات زیر توجه نمایید:

- تمرین بدون گزارش فاقد ارزش می‌باشد و نمره‌ای به آن تعلق **نمی‌یابد**.
- کامنت گذاری کدها در حد **لازم و کافی** الزامی می‌باشد.
- گذاشتن عنوان[۴] برای نمودارها و برچسب‌گذاری[۵] محورهای نمودار **الزامی** می‌باشد.
- در پیاده‌سازی بخش‌های مختلف، امکان استفاده از کتابخانه‌های آماده مرتبط با الگوریتم‌های یادگیری ماشین را به طور کلی ندارید. **موارد مجاز در صورت سوال بخش‌ها ذکر شده است.**
- برای نمایش نمودارها و عملیات ماتریسی می‌توانید از کتابخانه‌های numpy و matplotlib استفاده کنید. هم‌چنین برای خواندن داده‌ها به عنوان ورودی می‌توانید از pandas استفاده کنید.
- برای بهبود سرعت برنامه توصیه می‌شود که تا حد ممکن از عملیات ماتریسی استفاده شود.
- برای محاسبه معیارهای ارزیابی مانند دقت و ماتریس پریشانی می‌توانید از کتابخانه استفاده نمایید. **درصورت پیاده‌سازی آن‌ها نمره تشویقی تعلق خواهد گرفت. (۵+)**
- تا حد ممکن سعی کنید اصول لازم برای گزارش مهندسی را رعایت نمایید. **در این راستا برای گزارش‌هایی با ظاهر عالی نمره تشویقی در نظر گرفته شده است. (۵+)**
- مطابق قوانین دانشگاه هرگونه کپی‌برداری ممنوع می‌باشد و در صورت مشاهده نمره هر **دو طرف صفر** در نظر گرفته می‌شود.

تمرینات برنامه‌نویسی: (۱۰+۷۰٪)

۷- پیاده سازی دسته بند بیز ساده

الف) ابتدا مجموعه داده مربوطه را از مرجع [۳] دریافت کنید. پس از انجام پیش پردازش‌های لازم، دسته بند بیز ساده گاوسی را پیاده سازی کنید. به منظور گزارش دقت دسته بند از روش 6-fold cross validation استفاده کنید (توجه نمایید که در صورت کوچک بودن احتمالات می‌توانید از لگارتم احتمالات استفاده نمایید).

ب) قسمتی از داده‌ها را به عنوان داده تست در نظر گرفته و نمودار ROC را برای مدل آموزش داده شده رسم کرده و آن را تحلیل کنید.

۸- لاجستیک رگرسیون

مجموعه داده ارقام دست نویس MNIST را از مرجع [۴] دریافت کنید. با استفاده از روش One-vs-All داده‌ها را دسته بندی کنید. برای این منظور می‌توانید از رگرسیون لاجستیک خطی یا غیر خطی استفاده کنید. در صورت استفاده از مدل غیرخطی درجه آن را به صورت دلخواه انتخاب کنید. در این تمرین نیازی به پیاده سازی رگرسیون لاجستیک نیست. می‌توانید از کتابخانه آماده مانند sklearn استفاده کنید. اما توجه داشته باشید که قسمت One-vs-All باید پیاده سازی شود و از پارامتر multi_class استفاده نشود.

---

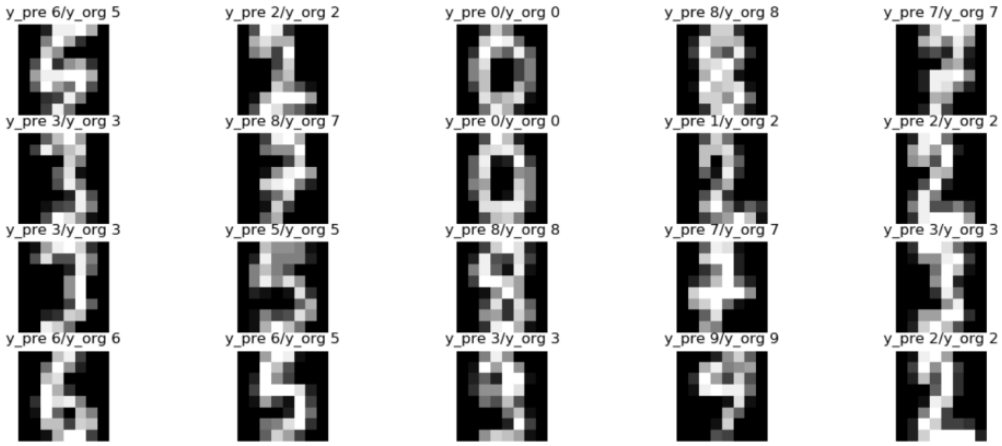
[4] Title
[5] Labeling

الف) پس از آموزش دسته بند، خطاهای مجموعه آموزش و تست و ماتریس پریشانی را گزارش کنید.

ب) ۲۵ داده از مجموعه تست به صورت تصادفی انتخاب کرده و برای هر داده کلاس واقعی و کلاس پیش‌بینی شده توسط مدل آموزش داده شده در تصویری مانند تصویر ۵ گزارش کنید.

ج) عملکرد این روش را با روش K نزدیک ترین همسایه که در تمرین سری دوم بررسی شد مقایسه کنید.

د) یکی از مشکلاتی که می‌تواند در استفاده از روش one-vs-all رخ دهد، مشکل یادگیری نامتوازن است. این مسئله را به صورت مختصر توضیح دهید و بیان کنید برای حل این مشکل چه پیشنهادی دارید.



تصویر ۵

مراجع:

[1] Ng, A. Y., & Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems* (pp. 841-848).

[2] https://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf

[3] https://archive.ics.uci.edu/ml/datasets/Wine

[4] http://yann.lecun.com/exdb/mnist/

موفق باشید