به نام خدا

دانشگاه صنعتی امیرکبیر

دانشکده‌ی مهندسی کامپیوتر

تمرین سری سوم یادگیری ماشین

دکتر احسان ناظرفرد

طراح سوال:

سید اردلان قریشی

محمدرضا امامی ناصری

دی ۱۳۹۹

توضیحات مهم:

- تمامی مستندات خود شامل گزارش و کدهای خود را در یک فایل فشرده با فرمت zip ذخیره کرده و با عنوان #StudentId_HW3.zip بارگذاری نمایید (به عنوان مثال 99131000_HW3.zip).
- مهلت انجام تمرین تا **ساعت ۲۳:۵۵** روز **یکشنبه** مورخ **۱۴ دی** می‌باشد و به هیچ وجه تمدید نمی‌شود.
- تمرین بدون گزارش فاقد ارزش می‌باشد و **نمره‌ای به آن تعلق نمی‌یابد**.
- تا حد ممکن سعی کنید اصول لازم برای گزارش مهندسی را رعایت نمایید (به بهترین گزارش نمره تشویقی تعلق می‌گیرد).
- مطابق قوانین دانشگاه هرگونه کپی برداری **ممنوع** می‌باشد و در صورت مشاهده نمره‌ی **هر دو طرف** صفر در نظر گرفته می‌شود.
- شما مجاز هستید برای تمامی تمرین‌ها **۷ روز در کل** و **با سقف حداکثر ۳ روز برای هر تمرین**، تاخیر بدون کسر نمره داشته باشید. به ازای هر روز تاخیر بیشتر، ۱۰٪ از نمره‌ی تمرین مربوطه کسر می‌شود.
- در صورت داشتن هرگونه ابهام می‌توانید از طریق ایمیل زیر سوال خود را مطرح نمایید:

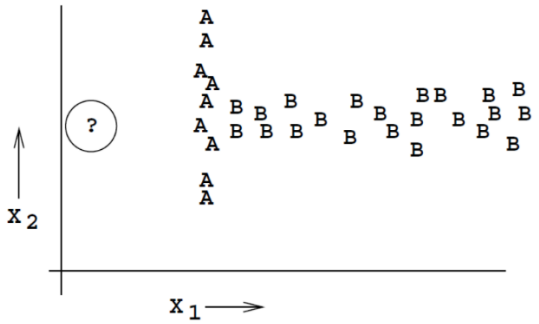
MLAUTFALL99@gmail.com

سوالات تشریحی

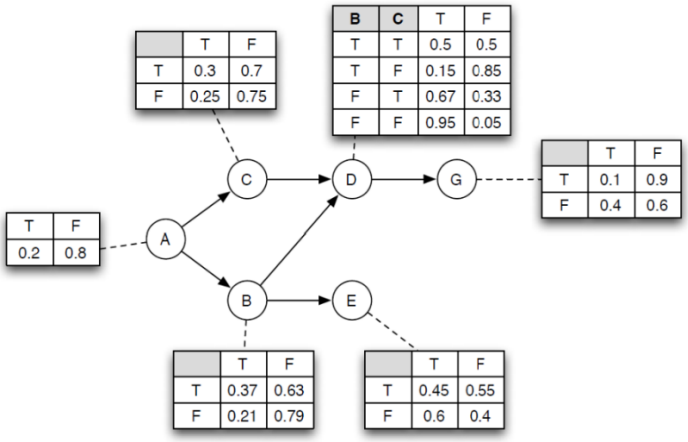
۱- توضیح دهید که عمل smoothing در بیز ساده[1] چیست و به چه منظور انجام می‌پذیرد؟

۲- با استفاده از مراجع ۱ و ۲ و سایر مراجع، دسته‌بندهای بیز ساده و رگرسیون لاجستیک[2] را با هم مقایسه کنید. (حداکثر در یک صفحه)

۳- برای هر کدام از داده‌های زیر که دارای دو ویژگی با مقدار حقیقی X_1 و X_2 هستند یک دسته‌بند بیز ساده گاوسی[3] را آموزش داده‌ایم. با ذکر دلیل تعیین کنید که برچسب داده تست (که با علامت سوال مشخص شده است) چه خواهد بود؟



۴- احتمال $P(B \mid D = T)$ را در شبکه بیزین[4] زیر حساب کنید.



[1] Naive Bayes
[2] Logistic Regression
[3] Gaussian Naive Bayes
[4] Bayesian Network

۵- نحوه انتخاب نقطه cut-off در یک مدل رگرسیون لاجستیک را شرح دهید.

۶- نسبت بخت[5] چیست؟ شرح دهید و نحوه‌ی استفاده آن را در رگرسیون لاجستیک بیان کنید.

۷- داده‌های آموزشی زیر که مربوط به افراد مختلفی است را در اختیار داریم. ستون Buy مشخص می‌کند که آیا فرد مورد نظر یک جنس مشخص (مثلاً کامپیوتر) را خریداری می‌کند یا خیر. با استفاده از دسته‌بند بیز ساده مشخص کنید که آیا افراد با مشخصات زیر، جنس مورد نظر را خریداری می‌کنند یا خیر؟

X_1 = (age = youth, income = high, student = yes, credit = fair)
X_2 = (age = senior, income = low, student = no, credit = excellent)
X_3 = (age = middle-aged, income = medium, student = no, credit = fair)

age	income	student	credit	Buy
youth	high	no	fair	-
youth	high	no	excellent	-
middle	high	no	fair	+
senior	medium	no	fair	+
senior	low	yes	fair	+
senior	low	yes	excellent	-
middle	low	yes	excellent	+
youth	medium	no	fair	-
youth	low	yes	fair	+
senior	medium	yes	fair	+
youth	medium	yes	excellent	+
middle	medium	no	excellent	+
middle	high	yes	fair	+
senior	medium	no	excellent	-

[5] Odds Ratio

سوالات پیاده‌سازی

توضیحات مهم:

- در روند اجرا انتخاب مقادیر برای تقسیم داده‌ها به مجموعه آموزش، ارزیابی و... به عهده دانشجو می‌باشد.
- حتما پارامترهای انتخاب شده برای برنامه خود و هرگونه شرایطی که درنظر گرفته‌اید را در گزارش خود بیاورید.
- برای بهبود سرعت برنامه توصیه می‌شود از عملیات ماتریسی استفاده کنید.
- در هر مرحله، نتایج خود را تحلیل کنید.
- کدهای خود را برای خوانایی بیشتر **کامنت گذاری** کنید.
- در تمامی سوال‌ها تنها مجاز به استفاده از کتابخانه‌هایnumpy، matplotlib و pandas می‌باشید.
- در پیاده‌سازی بخش‌های مختلف، امکان استفاده از کتابخانه‌های آماده مرتبط با الگوریتم‌های یادگیری ماشین را به طور کلی ندارید. **موارد مجاز در صورت سوال ذکر شده است.**
- گذاشتن عنوان برای نمودارها و برچسب گذاری محورهای نمودار الزامی می‌باشد.

۱- مجموعه داده‌ی Car Evaluation را از آدرس زیر دانلود کنید:

https://archive.ics.uci.edu/ml/datasets/Car+Evaluation

این مجموعه داده را در برنامه‌ی خود بارگزاری کرده و به دو قسمت آموزش و آزمون تقسیم کنید. (در صورتی که به پیش پردازش نیاز است، انجام داده و در گزارش خود بیاورید.)

تابعی بنویسید که دسته‌بند بیز ساده را با دریافت داده‌ها و پارامتر smoothing –که نشانگر فعال/غیرفعال بودن عمل smoothing است– اجرا کند.

نکته: در صورت کوچک بودن احتمالات می‌توانید از لگاریتم احتمالات استفاده نمایید.

الف) با استفاده از تابع بالا یک دسته‌بند بدون استفاده از smoothing ایجاد کنید.

خروجی مورد نظر: ماتریس درهم‌ریختگی و مقادیر sensitivity، specificity، false negative و false positive برای هر دو مجموعه‌ی **آموزش و آزمون**.

ب) با استفاده از تابع بالا یک دسته‌بند با استفاده از smoothing ایجاد کنید.

خروجی مورد نظر: ماتریس درهم‌ریختگی و مقادیر sensitivity، specificity، false negative و false positive برای هر دو مجموعه‌ی **آموزش و آزمون**.

ج) برای مجموعه‌ی آزمون نمودار ROC[6] را برای مدل آموزش داده شده‌ی قسمت الف رسم کنید و نتیجه را تحلیل کنید.

خروجی مورد نظر: نمودار ROC به همراه تحلیل آن.

۲- مجموعه داده‌ی MNIST را از آدرس زیر دانلود کنید:

http://yann.lecun.com/exdb/mnist/

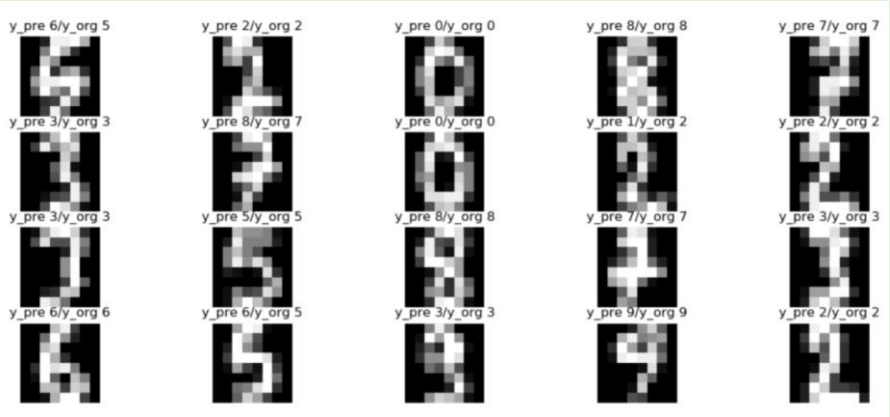
برنامه‌ای بنویسید که با استفاده از روش One-vs-All داده‌ها را دسته‌بندی کند. برای این منظور می‌توانید از رگرسیون لاجستیک خطی یا غیرخطی (با درجه‌ی دلخواه) موجود در کتابخانه‌ی آماده استفاده نمایید. توجه داشته باشید که شما باید بخش One-vs-All را خودتان پیاده‌سازی کنید؛ در نتیجه مجاز به استفاده از آرگومان multi_class='multinomial' موجود در کتابخانه‌ها نیستید.

الف) دسته‌بند را آموزش دهید و خطای آموزش و آزمون را به همراه ماتریش درهم‌ریختگی گزارش کنید.

خروجی مورد نظر: خطای آموزش و آزمون و ماتریس درهم‌ریختگی.

ب) ۲۵ داده از مجموعه‌ی تست به صورت تصادفی انتخاب کرده و برای هر داده، کلاس واقعی[7] و کلاس پیش‌بینی شده[8] توسط مدل آموزش داده شده را در تصویری گزارش کنید.

خروجی مورد نظر: تصویری مانند شکل زیر:



[6] Receiver operating characteristic
[7] Original Class
[8] Predicted Class

ج) عملکرد این روش را با روش K-نزدیک‌ترین همسایه[9] مقایسه کرده و توضیح دهید هر کدام از این روش‌ها برای چه مجموعه داده‌ای مناسب‌تر است.

با آرزوی موفقیت!

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

[9] K Nearest Neighbor (KNN)