یادگیری ماشین برای بیوانفورماتیک

نیم‌سال دوم ۹۹

مدرس: دکتر سلیمانی

دانشکده‌ی مهندسی کامپیوتر

تمرین سری دوم

۱. دسته‌بندی احتمالاتی (۳۰ + ۵ نمره)

۱ـ۱) (۱۵ + ۵ نمره) مسئله‌ی دسته بندی دو دسته‌ای را در نظر بگیرید که داده‌ها متعلق به یک فضای ویژگی دو بعدی بوده و احتمال شرطی دسته‌ها از توزیع گاوسی با ماتریس کوواریانس و میانگین‌های متفاوت بیایند. $(p(x|C_i) = N(\mu_i, \Sigma_i))$ همچنین احتمال پیشین روی دسته اول با برابر با π و روی دسته‌ی دوم را برابر با $1 - \pi$ در نظر بگیرید. $(p(C_1) = \pi, p(C_2) = 1 - \pi)$

الف) (۲ نمره) مرز دسته بند احتمالی بیز را بر حسب پارامترها و بردار ورودی x بیابید.

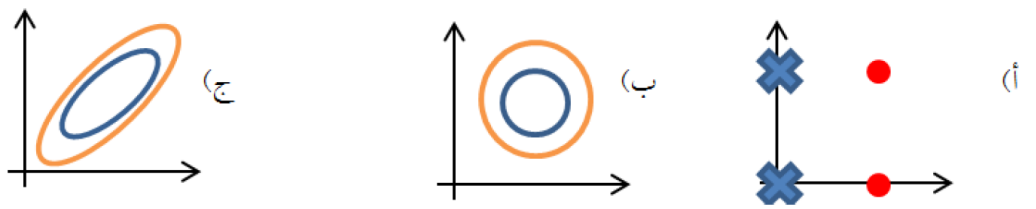
ب) (۵ نمره) در صورتی که ماتریس کوواریانس دو دسته یکسان در نظر گرفته شود، نشان دهید مرز به صورت یک ابرصفحه در می‌آید. پارامترهای این ابر صفحه را بر حسب پارامترهای π ، بردار میانگین و ماتریس کوواریانس دسته‌ها بنویسید.

ج) (۵ نمره امتیازی) در قسمت ب، فرض کنید تابع هزینه به این صورت تعریف شود که اگر داده کلاس اول اشتباهاً به داده کلاس دوم و داده کلاس دوم اشتباهاً به داده کلاس اول نسبت داده شود به‌ترتیب هزینه‌های L_1 و L_2 را دربرگیرد $(L_1, L_2 > 0)$، در خصوص تغییر مرز دسته بند با توجه به مقادیر مختلف L_1 و L_2 توضیح دهید.

د) (۴ نمره) در حالت $\Sigma_1 = \Sigma_2 = \sigma^2 I$ نشان دهید که دسته بند بیز برای هر داده، دسته‌ای را انتخاب می‌کند که میانگین آن به داده نزدیک تر است.

ه) (۴ نمره) برای قسمت ب، با تشکیل توزیع $P(x, y)$ و با استفاده از تخمین بیشینه درستنمایی و با فرض داشتن N داده آموزش $\{(x^{(i)}, y^{(i)})\}$ پارامترهای $\mu_1, \mu_2, \Sigma, \pi$ را به دست آورید. روابط قسمت الف چه تفاوتی با قسمت ب خواهند داشت؟ نیازی به نوشتن دقیق روابط قسمت الف نیست و توضیح تفاوت کافی است.

۱ـ۲) (۹ نمره) در شکل‌های زیر در حالت پیوسته منحنی‌های آبی و نارنجی، کانتورهای $P(x|C_1) = 0.5$ و $P(x|C_2) = 0.5$ را نشان می‌دهند (همچنین احتمال پیشین دسته‌ها در شکل‌های ب و ج برابر درنظر گرفته می‌شود) و در حالت گسسته احتمال نقاط هم‌رنگ با هم برابر و احتمال نقاط آبی دو برابر نقاط قرمز است. با استدلال بررسی نمایید که آیا فرض استقلال شرطی که در دسته‌بند Naive Bayes استفاده می‌شود، در هرکدام از این سه‌مورد برقرار است یا نه؟ چنان‌چه این شرط برقرار است مرز تقریبی که توسط این دسته‌بند برای دو دسته بدست می‌آید را در شکل زیر مشخص نمایید. (محور افقی ویژگی x_1 و محور عمودی ویژگی x_2 است)



۱ـ۳) (۶ نمره) یک مسئله logistic regression را درنظر بگیرید. اگر فیچرهای موجود در مجموعه‌های داده یکسان باشد، چه اتفاقی برای این نوع دسته‌بند خواهد افتاد؟ به عنوان مثال ۲ مجموعه داده زیر را درنظر بگیرید:

$$DS_۱ = (x^{(۱)}, y^{(۱)}), (x^{(۲)}, y^{(۲)}), ..., (x^{(N)}, y^{(N)})$$

$$DS_۲ = (x^{(۱)}, x^{(۱)}, y^{(۱)}), (x^{(۲)}, x^{(۲)}, y^{(۲)}), ..., (x^{(N)}, x^{(N)}, y^{(N)})$$

در صورت آموزش روی مجموعه داده دوم به‌جای اول، عملکرد این‌نوع دسته‌بند چه تفاوتی خواهد کرد؟

۲. دسته‌بند SVM (۱۵ نمره)

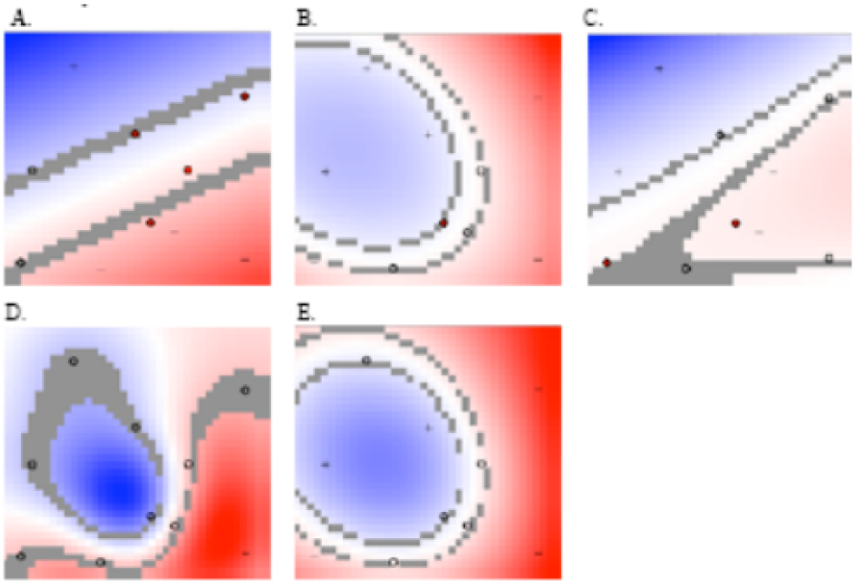
الف) (۴ نمره) درستی یا نادرستی گزاره‌های زیر را با توضیح مختصر مشخص نمایید.

- (۲ نمره) در SVM حاشیه‌ی نرم، با افزایش پارامتر C حاشیه‌ی جداسازی دو دسته افزایش می‌یابد و خطای آموزش کاهش می‌یابد.
- (۲ نمره) هسته (Kernel) گاوسی با پارامتر σ را درنظر بگیرید. در حالت کلی از بین مقادیر مختلف پارامتر σ برای هسته‌ی گاوسی، مقداری که منجر به حاشیه‌ی بزرگتر می‌شود، مقدار بهتری است.

ب) (۵ نمره) ثابت کنید در SVM خطی در حالتی که داده‌ها به‌صورت خطی جداپذیر هستند، عبارت زیر همواره برقرار است (دقت کنید که به صورت پیش فرض مقدار c=۱ فرض می‌شود. بنابراین در اصل در حال پیاده سازی به صورت soft margin هستید):

$$||w||^۲ = \sum_i \alpha_i$$

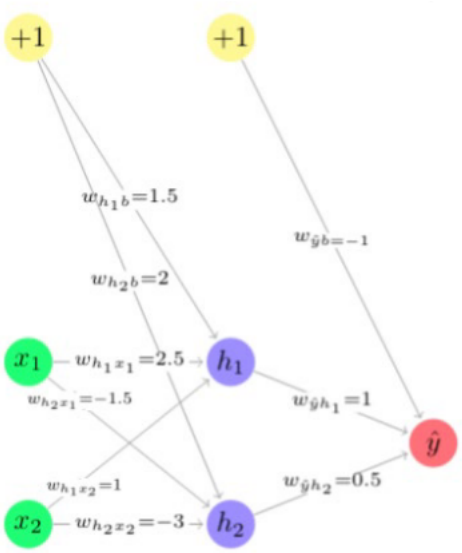
ج) (۶ نمره) برای یک مجموعه داده‌ی مشخص پنج دیاگرام حاصل از SVM با کرنل‌های مختلف در شکل زیر را درنظر بگیرید. با یک توضیح مختصر در جدول مربوطه مشخص کنید که هر دیاگرام می‌تواند توسط کدام‌یک از کرنل‌های زیر ایجاد شود.



RBF, $\sigma = ۰/۰۸$	
RBF, $\sigma = ۰/۵$	
RBF, $\sigma = ۲$	
Linear	
Second Order Polynomial	

۳. (۶ نمره) فرض کنید S مجموعه‌ی رشته‌هایی با طول حداکثر ۱۰۰ باشند که هر حرف در هر رشته از الفبای محدود A انتخاب شده باشد. برای هر $s \in \mathcal{S}$ که $s = a_1, \ldots, a_{100}$ داریم $a_j \in A$. کرنل $\mathcal{K} : \mathcal{S} \times \mathcal{S} \to \mathcal{R}$ را برای هر دو رشته در S به شکل $\mathcal{K}(s_1, s_2)$ نشان می‌دهیم و آن را تعداد زیر رشته‌های منحصر به فرد مشترک در s_1 و s_2 معرفی می‌کنیم. به طور مثال اگر فرض کنیم $A = \{a, e, i, o, u\}$ و $s_2 = aauuee$ و $s_1 = auue$ آنگاه خواهیم داشت: $\mathcal{K}(s_1, s_2) = 9$ چرا که زیر رشته های زیر مشترک هستند: $a, u, e, uu, ue, uue, au, auu, auue$ حالا اثبات کنید این تعریف یک تعریف معتبر برای تابع کرنل است.

۴. (۹ نمره) شکل زیر یک شبکه عصبی دو لایه با دو گره x_1 و x_2 در لایه ورودی، دو گره در لایه پنهان و یک گره در لایه خروجی را نشان می‌دهد. هر گره دارای یک ورودی بایوس با مقدار یک می باشد. فرض کنید از تابع سیگموئید به عنوان تابع فعالسازی در گره‌های لایه پنهان و خروجی استفاده می‌شود. سیگموئید تابعی است به فرم $g(z) = \frac{1}{1+e^{-z}}$ به طوری که $z = \sum_{i=1}^{n} w_i x_i$.(همچنین نرخ یادگیری را برابر با ۰/۱ درنظر بگیرید)



- (۳ نمره) فرض کنید x_1 و x_2 به‌ترتیب برابر با صفر و یک باشند. خروجی مقادیر گره‌های h_1, h_2 و $\hat{y}$ را بدست‌آورید
- (۶ نمره) فرض کنید x_1, x_2 و مقدار واقعی y به‌ترتیب برابر با صفر، یک و یک باشند. محاسبات مربوط به الگوریتم back-propagation را تنها برای یک گام انجام دهید.همچنین تابع لاس logistic regression را درنظر بگیرید .

بخش عملی

در این بخش برای پیاده‌سازی الگوریتم‌های زیر از مجموعه داده بیماران قلبی استفاده می‌کنیم که توصیه می‌شود پیش از پیاده‌سازی بخش‌های زیر، با مطالعه این لینک با ویژگی‌های مورد بررسی آشنا شوید. در تمام بخش‌ها اگر نیاز به داده اعتبارسنجی بود، با استفاده از ابزار موجود در scikit learn ۲۰ درصد مجموعه آموزش را برای این‌دسته اختصاص دهید. در فایل ژوپیتر هم چندین بخش ارزیابی وجود دارد که ورودی آن مجموعه داده‌ی تستی است که در اختیار شما قرار نگرفته‌است و در خروجی باید نتیجه ی ارزیابی متریک هایی که در آن قسمت از سوال آمده‌است با استفاده از داده‌ی ورودی چاپ شود.
در نهایت تاکید می‌شود تنها فایل ژوپیتری که پیوست‌شده است را تکمیل کنید و اگر نیاز به توضیحات اضافی بود، در انتهای آن بخش مربوطه ذکر کنید.

الف) Bayes, Naïve Bayes, Logistic Regression (۲۵ نمره)

در این سوال قصد داریم تا بیماران قلبی را از روی داده‌هایی که از پیش تهیه شده‌اند پیش بینی کنیم. در این داده، ویژگی‌هایی که بیشتر از ۵ عدد منحصر به فرد دارند را پیوسته و باقی ویژگی‌ها را گسسته در نظر بگیرید.

- (۵ نمره) در ابتدا ویژگی‌های گسسته را مستقل و ویژگی‌های پیوسته را دارای توزیع نرمال چند متغیری در نظر بگیرید. با این فرض دسته بند بیز را آموزش دهید و معیارهای F-score ، Accuracy ، Precision و Recall را بر روی داده‌های آموزش و تست گزارش کنید.
- (۵ نمره) حالا تمام ویژگی‌ها را مستقل فرض کنید و یک دسته بند Naïve Bayes آموزش دهید و معیار های ذکر شده را بر روی داده های تست گزارش کنید. (مجددا توزیع متغیرهای پیوسته را نرمال در نظر بگیرید.)
- (۵ نمره) با رسم نمودار توزیع نرمال بدست آمده برای هر یک از متغیرهای پیوسته در قسمت قبل، میزان اثر بخشی هر یک در دسته‌بند Naïve Bayes را بررسی کنید و با یکدیگر مقایسه کنید (در واقع باید برای هر متغیر پیوسته در یک نمودار توزیع نرمال برای هر دسته را رسم کنید).
- (۵ نمره) اگر فرض کنیم دسته‌بندی یک فرد ناسالم $(\text{target} == 1)$ در دسته‌ی سالم ها دو برابر هزینه‌ی بیشتری دارد، دسته‌بند Naïve Bayes را یکبار دیگر آموزش دهید و معیار Accuracy را با قسمت دوم مقایسه کنید.
- (۵ نمره) حالا سعی خواهیم کرد با استفاده از Logistic Regression ستون Target را تخمین بزنیم. برای اینکار از L2 Regularization استفاده کنید و با سنجیدن دقت روی داده‌ی اعتبارسنجی از بین مقادیر ۰/۰۱, ۰/۱, ۱, ۱۰ بهترین مقدار را برای lambda انتخاب و در نهایت معیار های گفته شده را بر روی داده‌ی تست گزارش کنید.

ب) SVM, Kernel (۱۲ نمره)

در این تمرین با استفاده از کتابخانه‌ی Scikit learn به دنبال تخمین ستون target هستیم.

- (۶ نمره) با فرض استفاده از کرنل خطی، میتوانید با تغییر دادن مقدار C بین حالات soft margin و hard margin حرکت کنید. در این حالت از بین چهار مقدار ۰/۱, ۰/۰۱, ۰/۰۰۱, ۰/۰۰۰۱ و با استفاده از معیار دقت روی داده‌ی اعتبارسنجی بهترین مقدار را انتخاب کنید و نتیجه‌ی نهایی را بر روی داده‌ی تست گزارش دهید. توضیح دهید چرا این مقدار بهترین نتیجه را دارد.
- (۶ نمره) با فرض استفاده از کرنل rbf با تغییر مقدار گاما به مقادیر ۰/۰۰۱, ۰/۰۰۰۱, ۰/۰۰۰۰۱ بهترین مقدار را با استفاده از معیار F-score و بر روی داده ی اعتبارسنجی بدست آورید و نتیجه‌ی نهایی را بر روی داده‌ی تست گزارش دهید. توضیح دهید بزرگ شدن مقدار گاما چه نتیجه ای ممکن است داشته باشد؟ (توجه داشته باشید که به صورت پیشفرض C برابر با ۱ فرض می‌شود. بنابراین در حال پیاده سازی soft margin هستید)

ج) kNN (۷ نمره)

در این تمرین هدف پیاده سازی الگوریتم kNN است. توابع موجود در فایل jupyter notebook را تکمیل کنید و سپس با استفاده از داده‌ی اعتبارسنجی تصمیم بگیرید مناسب ترین k از بین اعداد ۱، ۵، ۲۵، ۷۵ کدام است. در نهایت دقت نهایی را برای داده تست گزارش کنید.

د) ensemble Learning (۱۵ نمره)

در این سوال با کمک گرفتن از DecisionTreeClassifier در Sklearn قصد پیاده سازی Random forest و Adaboost را داریم. در اینجا توجه کنید که مجاز به کمک گرفتن از Sklearn به صورت مستقیم برای پیاده سازی این دو مورد نیستید.

- (۷ نمره) ابتدا Random forest را پیاده سازی می‌کنیم. برای آموزش هر درخت در جنگل، مقدار max-features را برابر با 'sqrt' قرار دهید. سپس توضیح دهید علت این تصمیم چیست؟ توجه کنید که باید از bootstrap sampling استفاده کنید. حالا تعداد درخت‌ها را از ۲۰ تا ۲۰۰ افزایش دهید. خطای آموزش و معتبرسازی را بر حسب تعداد درخت‌ها رسم کنید. (استفاده از کتابخانه برای bootstrap sampling منعی ندارد)

- (۸ نمره) حالا به دنبال adaboost می‌رویم. در اینجا هر عضو یک درخت تصمیم با عمق یک (decision stump) می‌باشد. همینطور توجه کنید که باید هر درخت را متناسب با وزن نقاط مختلف آموزش دهید. (از ورودی های درخت تصمیم در Sklearn استفاده کنید.) حالا تعداد دسته بند ها را افزایش دهید تا جایی که دیگر فایده ای نداشته باشد. توضیح دهید که در چه مرحله‌ای و به چه علت دیگر اضافه کردن درخت مفید نیست؟ به علاوه خطای آموزش هر دسته بند را محاسبه کرده و آن را رسم کنید. این نمودار نشانگر چیست؟

ه) neural network (۱۵ نمره)

در این بخش هدف پیاده‌سازی الگوریتم neural network با استفاده از فریم‌ورک keras می‌باشد. از این فرم‌ورک آشنایی با لایه Dense ، نحوه مدلسازی(Sequential) و نحوه اجرای فرآیندهای ترین و تست برای پیاده‌سازی ANN کفایت می‌کند.

- (۸ نمره) نحوه تنظیم هایپرپارامترها(مانند تعداد لایه‌ها و ...) برعهده خودتان است اما پارامترهای مختلفی را که تست‌کردید به همراه دقت آنها روی داده اعتبارسنجی را در انتهای بخش مربوط به شبکه‌های عصبی در فایل ژوپیتر بیاورید.

- (۷ نمره) در نهایت مدل با بالاترین دقت روی داده اعتبارسنجی را انتخاب کرده و نتیجه عملکرد آن روی داده تست را گزارش کنید.