به نام خدا

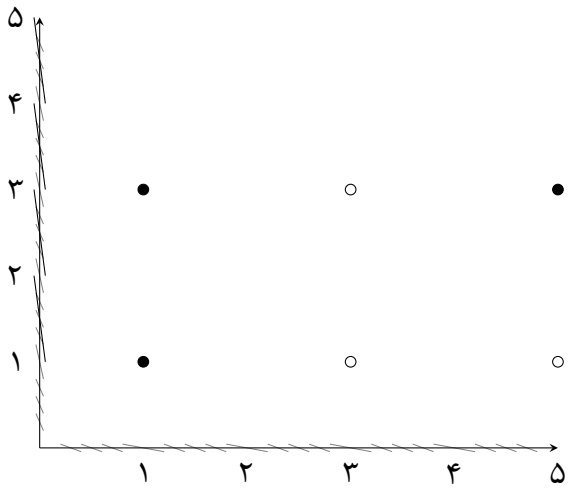
یادگیری ماشین — نیمسال دوم ۹۸-۹۷

تمرین سری سوم

موعد تحویل: ۱۸ فروردین — نمره: ۱۰۰ + ۲۰

سوال ۱: (30 Points) یادگیری جمعی

۱. داده‌های شکل زیر را در نظر بگیرید و با فرض این که هر دسته بند ضعیف یک خط افقی یا عمودی باشد، آدابوست را دو مرحله بر روی این داده‌ها اجرا کرده و وزن‌نمونه‌ها، مرز دسته‌بندی و خطای آموزش را در هر مرحله به دست آورید.



۲. فرض کنید با دو دسته‌بند ضعیف مواجه هستیم. دسته‌بند اول دسته‌بندی است که بر روی داده‌های آموزش $overfit$ می‌کند و قادر به ارائه عملکرد مثبت روی داده‌های دیده نشده نیست. همچنین دسته‌بند دوم هم با وجود قدرت تعمیم‌دهی بالا اما توانایی یادگیری مدل‌های پیچیده را ندارد. برای تقویت عملکرد هر کدام، از طریق روش‌های یادگیری جمعی، کدام از یک تکنیک‌های $bagging$ و یا $boosting$ را پیشنهاد می‌دهید؟ توضیح دهید.

۳. فرض کنید که $h_۱$ تا h_t و $\alpha_۱$ تا α_t فرضیه‌ها و ضرائب فرضیه‌های به دست آمده از اجرای آدابوست باشند و داشته باشیم

$$H_t(x) = \frac{۱}{۲}(\alpha_۱ h_۱(x) + \alpha_۲ h_۲(x) + ... + \alpha_t h_t(x))$$

و همچنین loss زیر را تعریف کنیم

$$Loss(y, H_t(x)) = \sum_{i=۱}^{N} e^{-y^{(i)} H_t(x^{(i)})}$$

- اولا نشان دهید عبارتی که در تکرار t از آدابوست بایستی کمینه شود عبارت زیر است

$$\sum_{i=۱}^{N} w_t^{(i)} e^{\frac{-۱}{۲}\alpha_t y^{(i)} h_t(x^{(i)})}$$

- ثانیا، نشان دهید که کمینه کردن عبارت بالا معادل است با کمینه کردن عبارت زیر

$$\sum_{i=۱}^{N} w_t^{(i)} * I(y^{(i)} \neq h_t(x^{(i)}))$$

- حال با مشتق گیری از $loss$ بر حسب α_t و صفر قرار دادن مشتق مقدار بهینه α_t که برابر با زیر است را به دست آورید:

$$\epsilon_t = \frac{\sum_{i=۱}^{N} w_t^{(i)} * I(y^{(i)} \neq h_t(x^{(i)}))}{\sum_{i=۱}^{N} w_t^{(i)}}$$

$$\alpha_t = ln(\frac{۱ - \epsilon_t}{\epsilon_t})$$

سوال ۲: (35 Points) شبکه پرسپترونی چندلایه

۱. یک شبکه عصبی را با k لایه در نظر بگیرید. اگر از تابع همانی به عنوان activation function در تمام لایه ها استفاده شود، در رابطه با خطی یا غیرخطی بودن مدل توضیح دهید.

۲. همانطور که میدانید تمام Boolean function ها را می توان با یک شبکه عصبی دو لایه مدل کرد:

- یک تابع بولین با N متغیر را در نظر بگیرید. بیشترین تعداد نورون ها در لایه مخفی چند است و در چه حالتی رخ می‌دهد.
- آیا می توان با افزایش لایه‌ها، تعداد نورون ها در لایه مخفی را نسبت به اندازه ورودی خطی کرد؟ با رسم شبکه عصبی نشان دهید.

۳. تعداد ۱۰ نمونه از دو کلاس C1 و C2 در زیر آمده است.

$$c۱ : [۰/۱, -۰/۲], [۰/۲, ۰/۱], [-۰/۱۵, ۰/۲], [۱/۱, ۰/۸], [۱/۲, ۱/۱]$$

$$c۲ : [۱/۱, -۰/۱], [۱/۲۵, ۰/۱۵], [۰/۹, ۰/۱], [۰/۱, ۱/۲], [۰/۲, ۰/۹]$$

- آیا این نقاط به صورت خطی تفکیک پذیر هستند؟ نشان دهید.
- اگر به صورت خطی تفکیک پذیر نیستند، یک شبکه عصبی با activation function پله برای آن طراحی کنید تا بتواند نقاط کلاسهای c1, c2 را به خوبی دسته بندی کند.

۴. نشان دهید اگر تابع هزینه در یک شبکه عصبی cross entropy باشد:

- خروجی شبکه عصبی که با استفاده از وزن های بهینه بدست می آید برابر تخمین $p(w_i|x)$ است.
- اگر $activation function$ برابر تابع سیگموید باشد، آنگاه:

$$\delta_i^L(i) = \frac{\partial \varepsilon(i)}{\partial v_j^L(i)}$$

$$\delta_j^L(i) = a(۱ - \hat{y}_j(i))y_j(i)$$

سوال ۳: (35 + 20 Points) تخمین ساختار دوم پروتئین (Protein Secondary Structure)

یکی از مسائل مطرح در حوزه بیوانفورماتیک، تخمین ساختار دوم پروتئین بر اساس ساختار اول آن یعنی رشته آمینواسیدی پروتئین است. برای توصیف ساختار دوم پروتئین می توان رشته آمینواسید آن را به زیررشته هایی تقسیم کرد و برای هر زیر رشته به صورت محلی ساختار دوم را بررسی کرد. در این مسئله از داده هایی با فرمت dssp استفاده می‌شود. برای آشنایی با فایل dssp می توانید به این لینک مراجعه کنید. هر فایل dssp مربوط به یک رشته پروتئین است. در این فایل که ساختار ماتریسی دارد، هر سطر مربوط به یک آمینواسید از پروتئین است که به ترتیب محل قرار گیری در رشته آمده است. هر چند سطر با هم به عنوان یک زیر رشته در نظر گرفته شده است. ستون های موجود در این ماتریس در لینک بالا توضیح داده شده است. ستون های ابتدایی مربوط به شماره آمینواسید و نام آن است و ستون‌های بعدی هر کدام یکی از ویژگی های ساختار دوم را مشخص می‌کند. در این تمرین از دو ستون این ماتریس استفاده می‌شود که برای راحتی کار شما، در هر فایل تنها همین دو ستون قرار دارند و ستون های دیگر حذف شده اند. لازم به ذکر است که داده‌هایی که در اختیار شما قرار گرفته اند مربوط به ۵۴۲۶ پروتئین است که هر کدام در فایل جداگانه ای قرار دارد. ستون اول در فایل‌های داده شده مربوط به نام آمینواسید است که به عنوان ویژگی در فرآیند آموزش استفاده می‌شود. ستون دوم مربوط به ساختار آن آمینو اسید است که با یک حرف مشخص شده است و به عنوان برچسب در فرآیند آموزش استفاده می‌شود. همانطور که در بالا توضیح داده شد، هر چند سطر از این ماتریس که برچسب یکسانی دارند، مربوط به یک زیررشته هستند. در این مسئله می‌خواهیم با داشتن زیررشته، ساختار دوم آن را تخمین بزنیم. ۰/۹ داده ها را برای آموزش و ۰/۱ را برای آزمون کنار بگذارید. در مورد نحوه‌ی انجام هر کدام از موارد زیر و مشاهداتتان از نتایج توضیح دهید.

۱. ابتدا در پیش پردازش زیررشته های به طول ۷ را نگه دارید و سایر زیررشته ها در نظر نگیرید.

۲. همانطور که با ساختار ویژگی ها آشنا شده اید می‌دانید که از نوع Categorical هستند و باید به نوع Numerical تبدیل شوند.

۳. با استفاده از پکیج sklearn و روش Random Forest دسته بندی را انجام دهید. در این روش hyperparameterهای عمق درخت و تعداد درخت وجود دارند که باید به درستی انتخاب شوند. مقادیر مناسب برای این پارامترها را انتخاب کنید و علت انتخاب خود را در گزارش بیاورید.

۴. پس از انتخاب مقادیر پارامترها، accuracy این روش را بروی داده های آموزش و آزمون گزارش کنید. همچنین precision و recall را برای داده‌های آموزش و آزمون برای هر دسته به صورت جداگانه محاسبه کنید. confusion matrix را برای داده‌های آموزش و آزمون به طور جداگانه محاسبه کرده و در گزارش بیاورید.

۵. با استفاده از پکیج sklearn و روش Multi Layer Perceptron دسته بندی را انجام دهید. همانند قسمت قبل accuracy دسته بندی، precision و recall برای هر دسته و همچنین confusion matrix را در گزارش بیاورید.

۶. (امتیازی) در بخش پیش پردازش برای یکسان سازی تعداد ویژگی ها برای هر نمونه، تنها زیررشته های به طول ۷ را نگه داشتیم و در نتیجه بخشی از داده را دور ریختیم. روشی ارائه دهید تا بتوانیم از زیررشته های با طول کمتر و بیشتر از ۷ هم استفاده کنیم. پس از تشکیل دوباره نمونه‌ها، بخش ۴ و ۵ را تکرار کنید.