به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

درس سیستم‌های هوشمند

تمرین شماره ۲

آبان ۱۴۰۰

فهرست سوالات

سوال ۱ : درخت تصمیم (تحلیلی)

جدول ۱ اطلاعات تعدادی بیمار را نشان می دهد:

جدول ۱-۱ : اطلاعات بیماران و وضعیت ابتلا به بیماری انسداد عروق برای دادگان آموزش

شماره	فشار خون	سطح کلسترول	مصرف سیگار	وزن	انسداد شرایین
۱	بله	نرمال	نه	اضافه وزن	بله
۲	نه	نرمال	بله	نرمال	نه
۳	نه	بحرانی	نه	اضافه وزن	بله
۴	نه	بالا	بله	اضافه وزن	بله
۵	بله	بحرانی	بله	چاق	بله
۶	بله	بالا	بله	نرمال	بله
۷	نه	بالا	نه	چاق	نه
۸	بله	نرمال	بله	نرمال	بله
۹	بله	بحرانی	نه	چاق	بله
۱۰	نه	نرمال	نه	اضافه وزن	نه
۱۱	نه	بحرانی	بله	نرمال	بله
۱۲	بله	بالا	نه	اضافه وزن	نه
۱۳	بله	نرمال	بله	اضافه وزن	بله
۱۴	بله	بالا	نه	چاق	نه

الف: طراحی طبقه‌بند

یک طبقه‌بند درخت تصمیم[۱] برای ویژگی ابتلا به بیماری انسداد عروق بر مبنای بهرهٔ اطلاعات[۲] را آموزش دهید.

[۱] Decision Tree

[۲] Information Gain

ب: آزمون طبقه‌بند

با استفاده از طبقه‌بند قسمت قبل طبقهٔ هر یک از نمونه‌های جدول ۲ را پیش‌بینی کرده و عملکرد مدل را به کمک ماتریس آشفتگی[۳] بررسی کنید.

جدول ۱-۲ : اطلاعات بیماران و وضعیت ابتلا به بیماری انسداد عروق برای دادگان آزمون

شماره	فشار خون	سطح کلسترول	مصرف سیگار	وزن	انسداد شرایین
۱۵	بله	نرمال	بله	چاق	بله
۱۶	بله	بالا	بله	چاق	بله
۱۷	بله	بالا	نه	نرمال	نه
۱۸	بله	نرمال	نه	نرمال	نه
۱۹	نه	نرمال	بله	اضافه وزن	بله

ج: افزایش قوام طبقه‌بند

چرا طبقه‌بندهای درخت تصمیم در برابر بیش‌برازش[۴] مقاوم[۵] نیستند؟ دو روش برای جلوگیری از این مشکل ارائه دهید.

---

[۳] Confusion Matrix

[۴] Overfitting

[۵] Robust

سوال ۲ : درخت تصمیم (شبیه‌سازی)

در این سوال با استفاده از پیاده‌سازی درخت تصمیم براساس الگوریتم ID3، قصد داریم داده های دادگان "prison_dataset.csv" را طبقه‌بندی کنیم. ویژگی هدف ما "نرخ بازگشت به زندان (تکرار جرم)[۶]" خواهد بود و میخواهیم براساس ویژگی‌های دیگر تصمیم گیری را انجام دهیم.

الف: طراحی طبقه‌بند

با نمونه برداری تصادفی[۷] و به صورت ۸۰ – ۲۰ از دادگان داده شده، آن را به داده‌های آموزش و آزمون[۸] تقسیم کنید. با استفاده از الگوریتم ID3 درخت خود را پیاده سازی کنید و آن را با داده های آموزش، آزمون دهید. معیار انتخاب ویژگی برتر را بهرهٔ اطلاعات در نظر گرفته و عمق درخت خود را ۳ در نظر بگیرید. در نهایت لازم است دقت طبقه بند برای داده های آزمایش و همچنین ماتریس آشفتگی را گزارش کنید. سپس عمق درخت را تغییر دهید و نتیجه گیری خود را براساس ماتریس آشفتگی توجیه کنید.

ب: استفاده از جنگل تصادفی

حال قصد داریم برای بهبود عملکرد طبقه بند، از الگوریتم جنگل تصادفی[۹] استفاده کنیم. بدین منظور می توانید داده ها و ویژگی ها را تقسیم کرده و تعداد K درخت (حداقل ۳ درخت) را آموزش دهید و در نهایت با استفاده از رای اکثریت[۱۰] دقت طبقه‌بند برای داده‌های آزمون و همچنین ماتریس آشفتگی را گزارش کنید. آیا دقت طبقه‌بند افزایش پیدا کرد؟ چرا؟

---

[۶] Recidivism - Return to Prison numeric

[۷] Random Sampling

[۸] Train and Test Data

[۹] Random Forest

[۱۰] Majority Voting

ج: استفاده از کتابخانه

در این قسمت با استفاده ازکتابخانه Scikit-Learn الگوریتم جنگل تصادفی را با درنظر گرفتن موارد زیر پیاده‌سازی کنید و دقت طبقه‌بند برای داده‌های آزمایش و همچنین ماتریس آشفتگی را گزارش کرده و آن را با قسمت ب سوال مقایسه کنید.

$$\begin{cases} \text{max\_depth} = 3 \\ \text{random\_state} = 0 \end{cases}$$

*** توجه ***: دقت کنید به دلیل اینکه جنگل تصادفی مربوط به کتابخانه Scikit-Learn مقادیر رشته[۱۱] برای ویژگی‌ها پشتیبانی نمی‌کند، لازم است که از "رمزگذار برچسب[۱۲]" برای این کار استفاده کنید. در این باره در اینترنت جست‌وجو کنید و با استفاده از کتابخانهٔ Scikit-Learn، از روش مناسب برای رفع مشکل استفاده کنید.

[۱۱] String

[۱۲] Label Encoder

سوال ۳: یادگیری بر اساس معیار

با توجه به دادگان "wine.csv" به سوالات زیر پاسخ دهید[۱۳]. ستون اول این داده کلاس هر داده را مشخص می‌کند و باقی ستون‌ها را به عنوان مجموعهٔ ویژگی‌ها در نظر بگیرید

*** توجه ***: در این سوال هیچ گونه پیش‌پردازشی بر روی داده مجاز نمی‌باشد. (همانند استانداردسازی، حذف ویژگی‌ها و ...)

الف: کا-همسایهٔ نزدیک

با پیاده سازی طبقه‌بند کا-همسایه نزدیک[۱۴]، عمل طبقه‌بندی را به ازای ۵ همسایه بر روی دادهٔ آزمون انجام دهید. برای این منظور ۲۰ درصد از داده را به آزمون و باقی را به دادهٔ آموزش اختصاص دهید. دقت و ماتریس آشفتگی را بر روی دادهٔ آزمون گزارش کنید. آیا می‌توان با تغییر تعداد همسایه مورد نظر، عملکرد این طبقه‌بند را بهبود بخشید؟

*** توجه *** : (در بخش فوق برای پیاده سازی طبفه‌بند کا-همسایهٔ نزدیک مجاز به استفاده از کتابخانه‌های آماده یادگیری ماشین نمی‌باشید.)

ب: یادگیری بر اساس معیار

در این قسمت ابتدا دو روش یادگیری بر اساس معیار[۱۵] (حاشیهٔ بزرگ همسایه‌های نزدیک[۱۶] و یک روش دیگر به دلخواه) را انتخاب نموده و ریاضیات و مسئله بهینه سازی مرتبط با آن را با دقت توضیح دهید. پس از اعمال آن ها بر روی داده، نتایج را با هم مقایسه کنید. حال طبقه‌بند ۵-همسایه نزدیک را برای همین دادگان استفاده کنید و نتایج را با قسمت قبلی مقایسه نمایید.

هم چنین در این حالت تعداد همسایه‌های مورد بررسی را تغییر داده (برای مثال ... ,۹ ,۵ ,۳ = K) و نتایج را بیان کنید. آیا می توانید روش سیستماتیکی برای تعیین تعداد همسایهٔ بهینه ارائه دهید؟ (با رسم نمودار توضیح دهید.)

*** توجه *** : در این بخش می توانید از کتابخانه metric-learn در پایتون و یا جعبه‌ابزارهای مرتبط با آن در MATLAB استفاده کنید.

---

[۱۳] برای اطلاعات بیشتر از ماهیت ستون‌ها می‌توانید به اینجا مراجعه کنید.

[۱۴] K Nearest Neighbor (KNN)

[۱۵] Metric Learning

[۱۶] Large Margin Nearest Neighbor (LMNN)

نکات:

- مهلت تحویل این تمرین، جمعه ۲۱ آبان است.
- انجام این تمرین به صورت یک نفره است.
- برای انجام تمرین‌ها فقط مجاز به استفاده از زبان های برنامه نویسی Python و MATLAB خواهید بود. در سوالاتی که از شما خواسته شده است یک الگوریتم را پیاده سازی کنید مجاز به استفاده از توابع آماده نمی‌باشید مگر اینکه در صورت سوال مجاز بودن استفاده از این توابع یا کتابخانه‌ها صریح ذکر شده باشد.
- کدهای مربوط به هر تمرین می‌بایست در پوشه‌ای با نام Codes در کنار گزارش کار شما موجود باشد. این کدها باید خوانا و به صورت مرتبط نام گذاری شده باشند، لذا توضیحات لازم را به صورت یادداشت[۱۷] در کدهای خود قرار دهید.
- گزارش شما در فرآیند تصحیح از اهمیت ویژه‌ای برخوردار است. لطفاً تمامی نکات و مفروضاتی که برای پیاده‌سازی‌ها و محاسبات خود در نظر می‌گیرید را در گزارش ذکر کنید.
- الزامی به ارائه توضیح جزئیات کد در گزارش نیست، ولیکن تحلیل و تفسیر نتایج بدست آمده الزامی است.
- گزارش‌ها تنها در قالب تهیه شده که روی صفحه درس در سامانه Elearn بارگذاری شده، تصحیح خواهند شد و به قالب‌های دیگر نمره‌ای تعلق نخواهد گرفت.
- در گزارش استفاده از زیرنویس برای تصاویر و بالانویس برای جداول الزامی است.
- در صورت مشاهدهٔ تقلب نمرات تمامی افراد شرکت‌کننده در آن **۱۰۰-** لحاظ می‌شود.
- لطفا گزارش ، فایل کدها و سایر ضمائم مورد نیاز را با ترتیب نام‌گذاری زیر در صفحه درس در سامانه یادگیری الکترونیکی بارگذاری نمائید.

HW[HW Number]_[LastName]_[StudentNumber].zip

- در صورت وجود هرگونه ابهام یا مشکل می‌توانید از طریق رایانامه‌های زیر با دستیاران آموزشی مربوطه در تماس باشید:

جناب آقای واهب – سوال ۱ - رایانامه { ovaheb@gmail.com }

جناب آقای ساعی‌زاده – سوال ۲ – رایانامه { alisaei90@gmail.com }

جناب آقای طلاکوب – سوال ۳ – رایانامه { rezatalakoob@yahoo.com }

[۱۷] comment