به نام خدا

یادگیری ماشین — نیمسال دوم ۹۸-۹۷

تمرین سری اول

موعد تحویل: ۳۰ بهمن — نمره: ۱۰۰

سوال ۱: Have Cold? (30 Points)

پرونده پزشکی مربوط به چند نفر که شامل اطلاعات دریافت شده هنگام مراجعه به پزشک و تشخیص پزشک درباره بیمار بودن یا نبودن آنهاست در جدول ۱ قابل مشاهده است. البته پزشک ما گاها برخی اطلاعات را به فرم دلخواه خود و بصورت کیفی وارد پرونده ها کرده است. بعنوان مثال او افراد بالای ۱۵ سال را بزرگسال و افراد زیر آن سن را جوان در نظر گرفته است. از دادگان آموزش برای ساختن درخت های تصمیم گیری استفاده نموده و سپس دقت دسته بند را روی داده های تست بررسی نمایید.

جدول ۱: مشخصات جمع آوری شده از بیماران، جدول آموزش

نام	جنسیت	سردرد	سرفه	سن	نژاد	سرماخوردگی دارد؟
پیتر	مرد	دارد	ندارد	۳۵	آمریکایی	بله
مدیسون	زن	دارد	ندارد	جوان	آسیایی	خیر
کلاوس	مرد	ندارد	ندارد	بزرگسال	آمریکایی	بله
ربکا	زن	دارد	ندارد	۱۵	اروپایی	خیر
ایزابلا	زن	ندارد	دارد	جوان	آسیایی	خیر
بارنی	مرد	دارد	دارد	بزرگسال	آسیایی	خیر
رونالد	مرد	ندارد	دارد	۶۵	آمریکایی	خیر
تد	مرد	دارد	ندارد	جوان	اروپایی	بله
تام	مرد	ندارد	دارد	جوان	آسایی	بله

جدول ۲: جدول تست

نام	جنسیت	سردرد	سرفه	سن	نژاد	سرماخوردگی دارد؟
راشل	مرد	ندارد	ندارد	۳۵	آمریکایی	بله
جیمی	زن	دارد	دارد	بزرگسال	اروپایی	خیر
لیزا	مرد	دارد	ندارد	۱۰	آمریکایی	خیر
ادوارد	مرد	ندارد	ندارد	۶۸	آسیایی	خیر
تونی	مرد	دارد	ندارد	۷۵	آسیایی	بله
جک	مرد	ندارد	ندارد	بزرگسال	آمریکایی	بله

۱. اگر ویژگی سردرد وجود نداشت؟

۲. اگر ویژگی نژاد وجود نداشت؟

۳. اگر پزشک برای تام تشخیص سرماخوردگی نمیداد؟

۴. اگر یک رکورد دیگر مانند جدول ۳ در جدول دادگان آموزش و تعداد D ویژگی اضافی با مقادیر تصادفی دارد/ندارد با احتمال برابر ($P(yes) = ۰/۵ = P(No)$) وجود میداشت؟

۵. نمودار توابع خطای مرحله آموزش و خطای تست را برای یک درخت تصمیم گیری، بر حسب حداکثر عمق محدود شده برای آن درخت رسم کنید.

۶. نمودار خطای آموزش و خطای تست را بر حسب سایز مجموعه آموزش رسم کنید.

جدول ۳: بخش چهارم سؤال اول

نام	جنسیت	سردرد	سرفه	سن	نژاد	سرماخوردگی دارد؟
تیلور	مرد	ندارد	بله	۱۳	اروپایی	خیر

سوال ۲: ML Measures (20 Points)

فرض کنید با یک بیماری سرطان مواجه هستیم و دو الگوریتم نیز در اختیار داریم که عملکردشان در مورد پیش‌بینی سرطان‌داشتن بیماران به شرح جداول زیر است:

جدول ۴: الگوریتم اول تشخیص سرطان

واقعیت / پیش بینی	مثبت	منفی
مثبت	۲۴	۲۷
منفی	۶	۹۴۳

جدول ۵: الگوریتم دوم تشخیص سرطان

واقعیت / پیش بینی	مثبت	منفی
مثبت	۲۸	۷۱
منفی	۲	۸۹۹

۱. معیارهای $accuracy$ و $precision$ و $recall$ را در مورد هر از یک الگوریتم‌های بالا محاسبه کنید.

۲. به طور شهودی از نظر شما کدامیک‌از متدهای بالا عملکرد بهتری در این وظیفه مشخص (تشخیص سرطان) از خود نشان می‌دهند؟

۳. از بین معیارهای $accuracy$ و $precision$ و $recall$ به نظر شما هر یک در چه دسته وظایفی معیارهای مناسبتری نسبت به بقیه معیارها هستند؟

سوال ۳: Deciding the Desease! - Practical (50 Points)

یک بیمارستان بین‌المللی در ژیلند برای پی بردن به علل به وجود آمدن بیماری ژنتیکی گیزووی از درخت تصمیم استفاده می‌کند. در این بیمارستان ۸۰۰ نمونه جمع‌آوری شده که از ۶۰۰ نمونه‌ی آن برای آموزش و ۲۰۰ نمونه‌ی باقی‌مانده برای آزمون استفاده می‌شود. در میان نمونه‌های جمع‌آوری شده هم نمونه‌های مربوط به افراد دارای بیماری و هم بدون بیمای وجود دارد. برای بررسی بیماری از ویژگی‌های نژاد، جنسیت و رخداد یا عدم رخداد ۲۰ نوع جهش مختلف در افراد استفاده شده است.

۱. به دلیل تازه‌کاری بعضی از پرسنل بیمارستان، ویژگی‌های افراد به صور مختلف درج شده‌اند. مثلا جنسیت در بعضی موارد به صورت F و در بعضی دیگر به صورت female درج شده است. لازم است پیش از شروع به تحلیل داده، داده‌ها پاک‌سازی و مقادیر یکسان‌سازی شوند. برای انجام این کار توابع زیر در پایتون می‌تواند برای شما کمک کننده باشد:

Numpy: unique, vectorize, apply
Pandas: read csv

گفتنی است کتابخانه‌های numpy و pandas کتابخانه‌های بسیار سریع و بهینه‌ای هستند که برای پردازش داده‌های بزرگ از آن‌ها استفاده می‌شود و قطعا در آینده راهتان به آن‌ها می‌افتد. بنابراین از همین الان توصیه می‌شود استفاده از آن‌ها را یاد بگیرید :).

۲. در گام بعدی هدف یاد گرفتن درخت تصمیمی است که بتواند افراد دارای بیماری را از سایرین جدا کند. برای انتخاب ویژگی‌های رئوس این درخت از معیار Information gain استفاده کنید. این معیار به صورت زیر محاسبه می‌شود:

$$
\begin{aligned}
&Entropy\ H(x) = -\sum_{x_i \in x} p(x_i) log(p(x_i)) \\
&Information\ Gain\ Gain(S, A) = H(S) - \sum_{S_v \in values(A)} \frac{|S_v|}{|S|} H(S_v)
\end{aligned}
\tag{۱}
$$

عمق درخت یکی از هایپرپارامترهای این دسته‌بند است. فعلا در این مرحله عمق را به صورت پارامتری نگه دارید تا در گام‌های بعدی بتوانید مقادیر مختلفی را روی آن امتحان کنید.

۳. درخت تصمیم خود را با محدودیت عمق ۱ (فقط یک ویژگی) تا ۲۲ (تمام ویژگی‌ها) آموزش دهید. (برای جلوگیری از ابهام لازم به ذکر است برای تک تک اعداد ۱ تا ۲۲ لازم است این کار را انجام دهید). دقت دسته‌بند را برای داده‌های آموزش و آزمون به ازای هر محدودیت عمق محاسبه کنید. نمودار دقت روی داده‌های آموزش و آزمون را بر حسب محدودیت عمق درخت (به عنوان معیاری از پیچیدگی مدل) رسم کنید. تمام افت و خیزهای نمودار را توصیف کنید.

۴. با استفاده از نمودار رسم شده در مرحله‌ی پیش، به نظرتان چه عمقی برای این درخت مناسب‌تر است؟ آیا می‌توان دقت دسته‌بند روی داده‌های آزمون را برای این عمق به عنوان معیار نهایی عملکرد مدل گزارش کرد؟

۵. به استفاده از 5fold-cross validation مناسب‌ترین عمق را برای درخت انتخاب کنید. سپس درخت را با این محدودیت عمق روی تمامی داده‌های آموزش تعلیم دهید. معیارهای sensitivity و specificity را برای داده‌های آزمون گزارش کنید. ارزش هر کدام از این معیارهای ارزیابی در مقابل معیار دقت چیست و چه زمانی هر کدامشان اهمیت بیشتری پیدا می‌کنند؟

۶. می‌توان برای کوچک کردن درخت، به جای اینکه شرط پایان را رسیدن به محدودیت عمق و یا خالص شدن یک برچسب در یک نود، درصد خلوص بیشتر از یک حد مشخص در نود گذاشت (برای مثال وقتی بیش از ۸۰٪ نمونه‌ها در یک نود درخت یک برچسب یکسان داشتند، به آن نود همان برچسب تعلق گیرد و دیگر گسترش داده نشود.). همین کار را برای محدودیت عمق ۲۲ (تمام ویژگی‌ها) انجام دهید. دقت به دست آمده را با دقت قبلی مقایسه کنید. از این اتفاق چه نتیجه‌ای میگیرید؟

۷. برای بخش قبلی دقت دو حالت هرس‌شده و هرس‌نشده را از طریق paired t-test مقایسه و نتایج تست را تفسیر کنید.