هوش مصنوعی **تمرین شماره‌ی ۳**

مدرس: **دکتر فدایی** طراحان: **نگار مرادی، علیرضا اربابی**

مهلت تحویل: دوشنبه ۸ خرداد ۱۴۰۲، ساعت ۲۳:۵۵

درخت تصمیم

فرخنده مدتی است که از دریافت ایمیل‌های spam کلافه شده است و هم‌اکنون پس از مطالعه مباحث مربوط به یادگیری ماشین، قصد دارد تا ایمیل‌های ورودی به آدرس خود را دسته‌بندی کرده و ایمیل‌های Spam را از ایمیل‌های Ham (ایمیل‌های غیر اسپم) جدا کند.

- جدول ۱-۱ شامل ۱۴ داده و شامل مجموعه دادگان آموزش شما می‌باشد.
- جدول ۱-۲ شامل ۶ داده و شامل مجموعه دادگان آزمون شما می‌باشد.

قسمت اول)

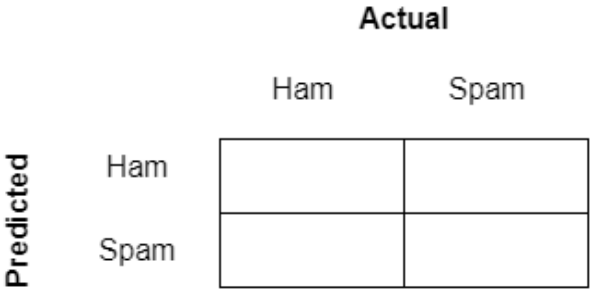
یک طبقه‌بند درخت تصمیم مبتنی بر Information gain را با عمق ۳ (با احتساب ریشه و برگ‌ها) برای پیش‌بینی spam و یا ham بودن ایمیل ها را بر روی مجموعه دادگان جدول ۱-۱ آموزش دهید. علاوه بر نشان دادن درخت تصمیم نهایی، مراحل محاسبات خود، برای ساخت آن را بنویسید.

قسمت دوم)

با استفاده از طبقه‌بند ساخته‌شده در قسمت اول، طبقه هر کدام از داده های آزمون جدول ۱-۲ را پیش‌بینی کنید.

قسمت سوم)

حال با طبقه‌بندی انجام شده برای دادگان آزمون، برای ارزیابی عملکرد طبقه‌بند، ابتدا ماتریس درهم‌ریختگی را ایجاد کرده و سپس Accuracy و مقادیر Precision و Recall را برای هر نتیجه محاسبه کنید. برای مطالعه بیشتر درباره این موضوع میتوانید به این لینک مراجعه کنید. ماتریس درهم‌ریختگی را با ساختار زیر نشان دهید:



قسمت چهارم)

تفاوت روش‌های یادگیری ماشین random forest و d-tree چیست؟ استفاده از روش random forest چه مزیتی برای ما دارد؟ تفاوت متریک های bias و variance را در هر دو روش بررسی کرده و علت تفاوت را شرح دهید.

شماره	نرخ شکایت از فرستنده	طول بدنه ایمیل	فرمت ایمیل	دامنه ایمیل	تشخیص
۱	پایین	متوسط	مشکوک	gmail	Ham
۲	بالا	کوتاه	مشکوک	gmail	Spam
۳	پایین	متوسط	مشکوک	گمنام	Ham
٤	پایین	کوتاه	مشکوک	gmail	Spam
٥	بالا	کوتاه	نامشکوک	academic	Ham
٦	پایین	طولانی	نامشکوک	gmail	Spam
۷	بالا	متوسط	نامشکوک	gmail	Spam

۸	بالا	کوتاه	مشکوک	academic	Spam
۹	پایین	طولانی	مشکوک	گمنام	Spam
۱۰	پایین	کوتاه	نامشکوک	gmail	Ham
۱۱	بالا	طولانی	نامشکوک	academic	Spam
۱۲	بالا	طولانی	مشکوک	گمنام	Spam
۱۳	بالا	متوسط	مشکوک	academic	Spam
۱٤	پایین	متوسط	نامشکوک	گمنام	Ham

جدول ۱-۱

شماره	نرخ شکایت از فرستنده	طول بدنه ایمیل	فرمت ایمیل	دامنه ایمیل	تشخیص
۱	پایین	کوتاه	مشکوک	کم	نامشکوک
۲	بالا	کوتاه	نامشکوک	متوسط	مشکوک
۳	بالا	کوتاه	مشکوک	زیاد	مشکوک
٤	بالا	متوسط	مشکوک	زیاد	مشکوک
٥	پایین	متوسط	مشکوک	کم	نامشکوک
٦	بالا	طولانی	نامشکوک	متوسط	مشکوک

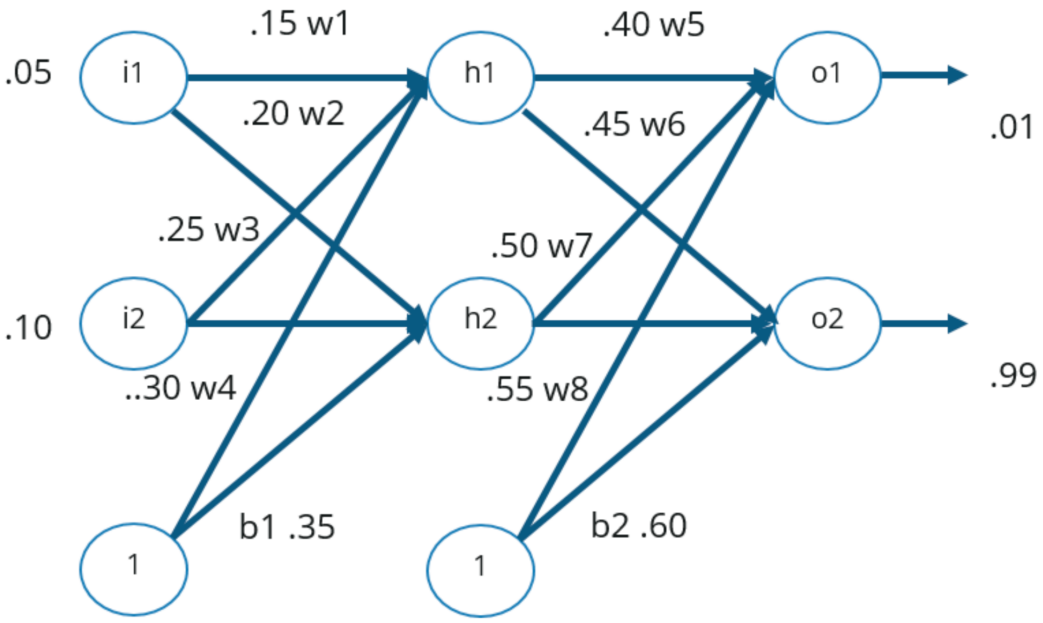
جدول ۱-۲

شبکه‌های عصبی

قسمت اول)

شبکه‌ی عصبی زیر را در نظر بگیرید. این شبکه یک input layer، یک hidden layer و یک output layer دارد. مقادیر bias و همچنین وزن اولیه برای هر نود داده شده است. activation function برای hidden layer و output layer از نوع sigmoid است.

$$Sigmoid(x) = \frac{1}{1 + e^{-x}}$$



۱. خروجی نودهای h_1, h_2, o_1, o_2 را به دست آورید.

۲. با فرض مقدار targetهای $t_1 = 0.01, t_2 = 0.99$ برای o_1, o_2 (خروجی واقعی) با استفاده از تابع خطای زیر مقدار جدید w_5 را محاسبه کنید. (learning rate را 0.5 در نظر بگیرید)

$$E_{total} = \sum \frac{1}{2}(target - output)^2$$

قسمت دوم)

۱. دو activation function معروف را نام ببرید و مزایای آن‌ها را بنویسید.