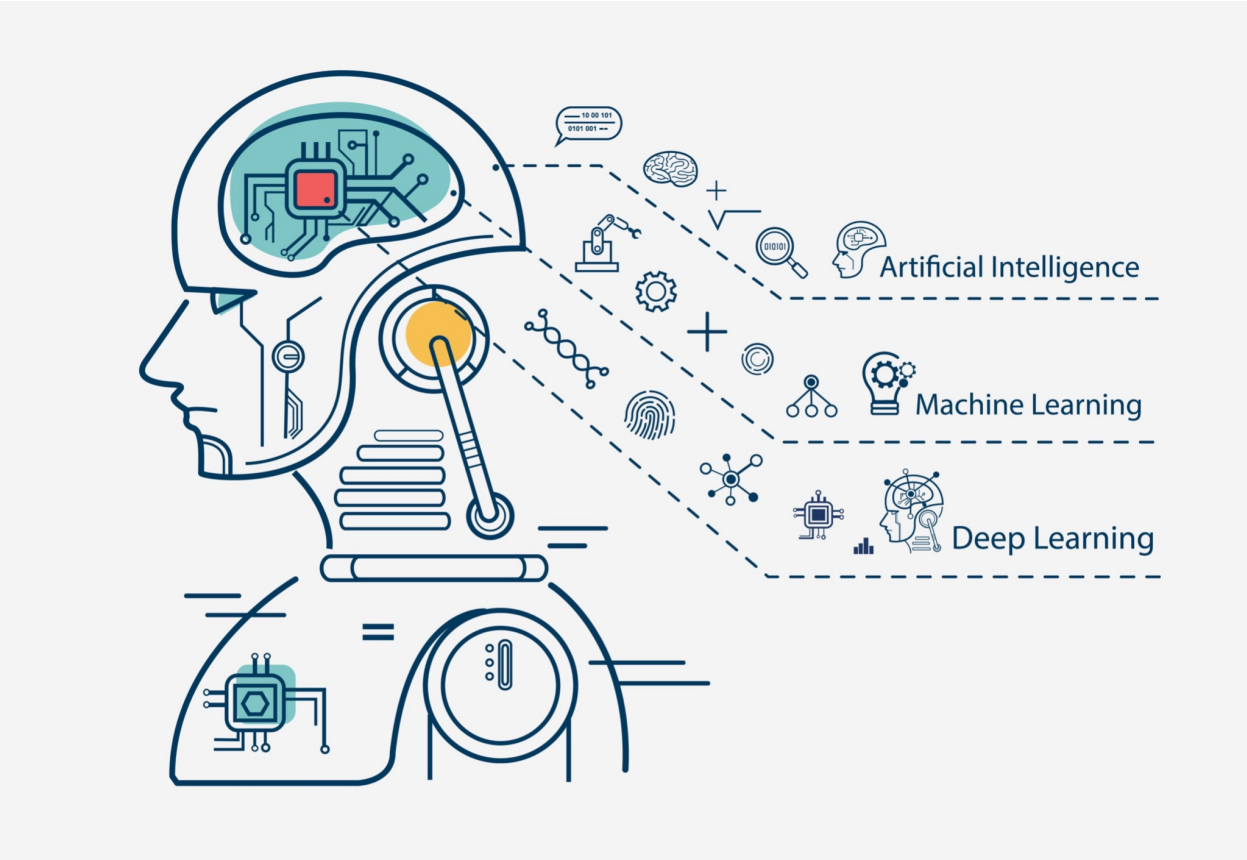


دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

تمرین سری سوم یادگیری ماشین

دانشکده مهندسی کامپیوتر

استاد درس: دکتر ناظرفرد

اردیبهشت ۱۴۰۰

- تمامی مستندات شامل گزارش به همراه کدهای خود را در یک فایل فشره با فرمت zip ذخیره کرده و با عنوان stdNum_HW2.zip که stdNum شماره دانشجویی شما است در سامانه بارگزاری کنید.
- سوالات ستاره‌دار(*) نمره اضافی داشته و انجام آن‌ها اجباری نمی‌باشد.
- مهلت انجام تمرین تا ساعت **۲۳:۵۵ روز شنبه ۸ خرداد** می‌باشد.

سوال‌های تشریحی(ماشین‌های بردار پشتیبان و مدل‌های ترکیبی)

سوال ۱) صحت هر یک از موارد زیر را بررسی کرده و دلایل خود را توضیح دهید.

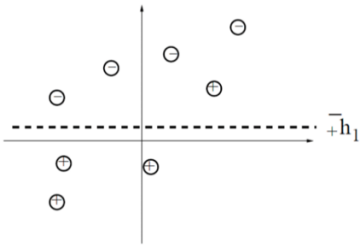
الف) ماشین‌های بردار پشتیبان[۱] پارامتریک‌اند.

ب) مقدار حاشیه بدست آمده برای دو ماشین بردار پشتیبان با کرنل‌های متفاوت که برای داده‌های یکسان آموزش دیده‌اند، می‌تواند معیاری برای میزان کارایی مدل باشد.

ج) ماشین‌های بردار پشتیبان همواره در برابر بیش‌برازش مقاوم می‌باشند.

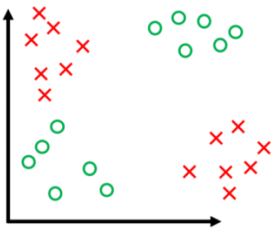
د) وجود داده‌های پرت و نویز بر روی ماشین‌های بردار پشتیبان بی‌تاثیر است.

سوال ۲) برای داده‌های زیر الگوریتم آدابوست را اجرا کرده و در تکرار اول مرز تصمیم h_1 بدست آمده است. وزن α_1 که توسط الگوریتم آدابوست به h_1 اختصاص داده می‌شود چقدر است؟(وزن‌های اولیه برابر $\frac{1}{8}$ می‌باشند.)



شکل ۱

سوال ۳) امکان استفاده از ماشین‌های بردار پشتیبان برای مسئله زیر را مورد بررسی قرار دهید. در صورت امکان راهکار خود را بطور کامل با ذکر مقادیر و جزییات شرح دهید.



شکل ۲

[۱] Support Vector Machine (SVM)

سوال ۴) سوالات زیر را در مورد روش‌های ترکیبی پاسخ دهید.

الف) آیا الگوریتم آدابوست در برابر داده‌ی نویز حساس است؟ چرا؟

ب) الگوریتم آدابوست با استفاده از هر نوع دسته‌بند ضعیف و یا ترکیب چند دسته‌بند ضعیف درنهایت به خطای صفر می‌رسد.

- کدهای خود را به زبان پایتون و ترجیحا در محیط jupyter پیاده‌سازی کنید. می‌توانید تحلیل خودتان را به عنوان سلول‌های متنی در همان محیط ارائه کنید.
- نظم در نوشتن گزارش و کدها می‌تواند به کسب نمره‌ی بهتر به شما کمک کند. برنامه نوشته شده خوانا و کامنت گذاری مناسب داشته باشد.
- در پیاده‌سازی بخش‌های مختلف، امکان استفاده از کتابخانه‌های آماده مرتبط با الگوریتم‌های یادگیری ماشین را به طور کلی ندارید. **موارد مجاز در صورت سوال بخش‌ها ذکر شده است.**
- برای نمایش نمودارها و عملیات ماتریسی می‌توانید از کتابخانه‌های numpy و matplotlib استفاده کنید. همچنین برای خواندن داده‌ها به عنوان ورودی می‌توانید از pandas استفاده کنید.
- برای محاسبه معیارهای ارزیابی مانند دقت و ماتریس درهم‌ریختگی می‌توانید از کتابخانه آماده استفاده کنید.
- مطابق قوانین دانشگاه هرگونه کپی‌برداری ممنوع می‌باشد و در صورت مشاهده نمره هر **دو طرف صفر** در نظر گرفته می‌شود.
- در صورت داشتن سوال می‌توانید با ایمیل تدریس یاران درس در تماس باشید:
  hse.khalilian08@gmail.com , hamid.dargahi0072@gmail.com

سوال‌های پیاده‌سازی

مسئله ۱) پیاده‌سازی ماشین بردار پشتیبان

دیتاست lsvt-voice-rehabilitation را از ادرس زیر دانلود کرده و به سوالات پاسخ دهید. برای ارزیابی از ۱۰-fold-cross-validation استفاده کنید. در این تمرین مجاز به استفاده از کتابخانه آماده هستید.

https://archive.ics.uci.edu/ml/datasets/LSVT+Voice+Rehabilitation

الف) داده‌ها را به روش های زیر دسته بندی کنید سپس به سوالات که در زیر آمده است پاسخ دهید.

- کرنل خطی
- کرنل چند جمله‌ای(پارامترهای d,r)

- کرنل RBF (پارامتر گاما)
- سیگموئید(پارامتر r)

ب) معیار دقت[۲] و F۱-measure را برای هر یک از دسته‌بندهای بالا بدست آورید(برای هر یک از پارامترهای یاد شده حداقل سه مقدار متفاوت در نظر بگیرید.)

ذ) تاثیر پارامترهای هر کرنل بر کارایی مدل‌ها را تحلیل کنید.

د) بهترین مدلی را که یافتید، مشخص نمایید.

مسئله ۲) بررسی عملکرد مدل‌های ترکیبی

توضیح: داده‌های مورد استفاده در این تمرین مربوط به تصاویر اعداد دست نوشته انگلیسی است. از هر تصویر، ۱۶ویژگی استخراج‌شده که این ویژگی‌ها به همراه برچسب تصویر متناظر، در اختیار شما قرار داده شده است. هدف، طراحی دسته‌بند مناسب برای این داده‌ها است. لازم به ذکر است که داده های یادگیری و تست به صورت مجزا و به ترتیب در فایل data_train.csv و data_test.csv قرار دارد.

الف) با استفاده از روش Random Forest یک دسته‌بند مناسب طراحی کنید. دسته‌بندی را با استفاده از ترکیب نتیجه‌ی ۱۵ درخت تصمیم انجام دهید. هر درخت تصمیم را نیز با استفاده از۳ ویژگی که به صورت تصادفی انتخاب می‌شود، آموزش دهید. برای آموزش درخت‌های تصمیم می‌توانید از کتابخانه‌ی scikit-learn استفاده کنید. بیشینه‌ی عمق درخت‌ها را نیز ۳ در نظر بگیرید. دقت دسته‌بندی و همچنین ماتریس درهم‌ریختگی را برای داده های تست در گزارش خود ارائه کنید.

ب) با استفاده از روش AdaBoost یک دسته‌بند مناسب طراحی کنید. دسته بندی را با استفاده از ترکیب نتیجه‌ی ۱۰درخت تصمیم انجام دهید. برای آموزش درخت‌های تصمیم همچون قبل می‌توانید از کتابخانه‌ی scikit-learn استفاده کنید. دقت دسته‌بندی را برای داده های تست در گزارش خود ارائه کنید.

ج) با استفاده از روش AdaBoostقسمت قبل را با ۵، ۲۰ و۵۰ دسته‌بند درخت تصمیم تکرار کرده و برای هر کدام دقت دسته‌بندی را برای داده های تست در گزارش خود ارائه کنید.

***د)** در چند سال اخیر روش هایGradient Boosting بسیار مورد توجه قرار گرفته اند. یکی از کتابخانه هایی که برای این روش ها منتشر شده است، کتابخانه‌ی XGBoost است. با بررسی این کتابخانه، یک دسته‌بند مناسب برای داده های این تمرین طراحی کنید. انتخاب پارامتر های بهینه در این قسمت به عهده شما است. پس از پایان یادگیری، دقت دسته بندی را بر روی داده های تست برای دسته‌بند پیشنهادی خود در گزارش ارائه کنید.

[۲] Accuracy