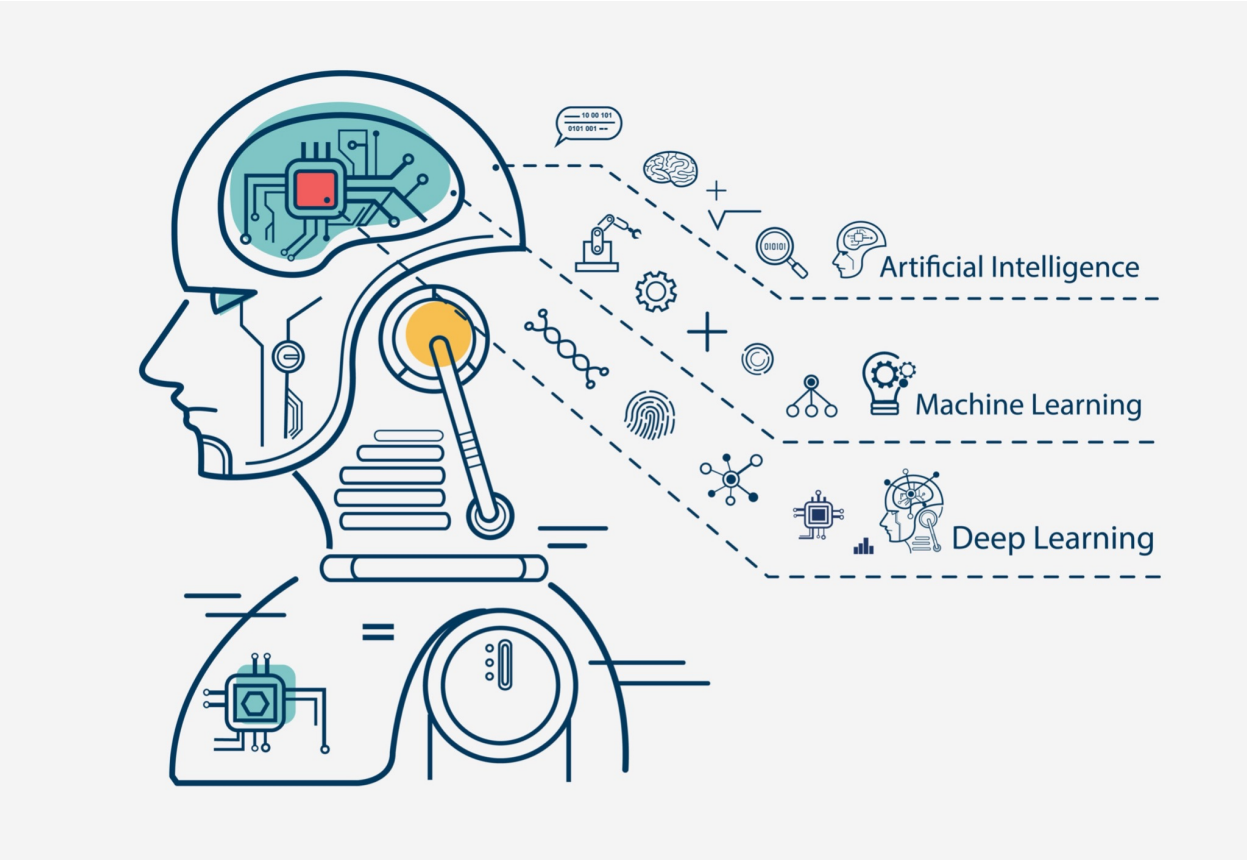


دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

تمرین سری دوم یادگیری ماشین

دانشکده مهندسی کامپیوتر

استاد درس: دکتر ناظرفرد

فروردین 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 ۱۸ اردیبهشت** می‌باشد.

سوال‌های تشریحی

سوال ۱) هرس درخت تصمیم چه تاثیری بر بیش‌برازش دارد؟ این هرس چه زمانی باید انجام شود؟ توضیح دهید. راهکارهای دیگری برای جلوگیری از بیش‌برازش را در درخت تصمیم توضیح دهید.

سوال ۲) ابزار وکا[1] را دانلود کنید. با توجه به مجموع داده‌ی labor از مجموعه داده‌های نمونه وکا به موارد زیر پاسخ دهید.

الف) مجموع داده مورد نظر را از تب preprocess بارگذاری کرده و از تب classify درخت تصمیم J۴۸ را با تنظیمات پیش‌فرض و با ۱۰-fold-cross-validation آموزش داده و دقت ماتریس درهم ریختگی[2] آن را گزارش کنید. سپس از روی ماتریس درهم‌ریختگی مقادیر، TP، TN، FP،FN، Precision، Recall، F۱-Measure را برحسب خانه‌های ماتریس درهم ریختگی بدست آورید. درخت تصمیم ساخته شده را رسم کنید. داده‌های زیر در کدام کلاس قرار می‌گیرد؟ مراحل یافتن کلاس این داده را با داشتن درخت تصمیم توضیح دهید.

feature	value	feature	value
duration	1	shift-differential	20
wage-increase-first-year	3	education-allowance	yes
wage-increase-second-year	6	statutory-holidays	12
wage-increase-third-year	4	vacation	generous
cost-of-living-adjustment	tcf	longterm-disability-assistance	yes
working-hours	35	contribution-to-dental-plan	full
pension	ret_allw	bereavement-assistance	no
standby-pay	11	contribution-to-health-plan	half

شکل ۱

ب) پارامتر unpruned درخت تصمیم چه چیزی را کنترل می‌کند؟ این پارامتر را از مقدار پیش‌فرض False به True تغییر داده و تمام موارد خواسته شده در قسمت قبل را انجام داده و گزارش کنید. تفاوت درخت آموزش داده شده در این بخش نسبت به بخش قبل چیست؟

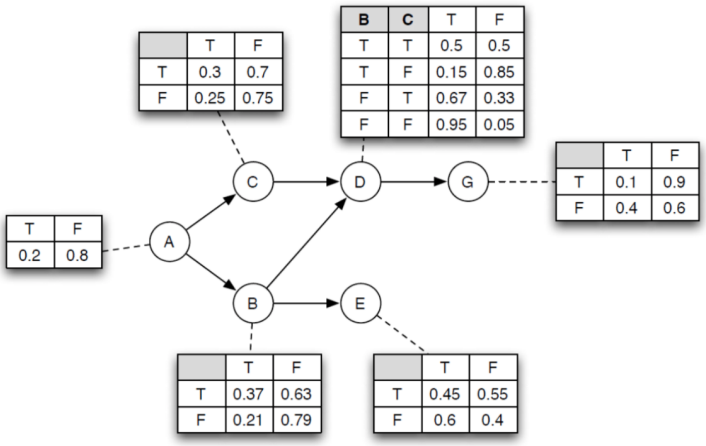
[1] Weka
[2] Confusion Matrix

سوال ۳) برای دستیابی به کارایی خوب در الگوریتم KNN، در صورت افزایش ابعاد داده‌ها، اندازه داده‌های مورد نیاز تغییر می‌یابد؟ صحت پاسخ ارایه شده را با دلایل کافی مورد بررسی قرار دهید.

سوال ۴) تفاوت مدل‌های Generative و Discriminative را بیان کنید.

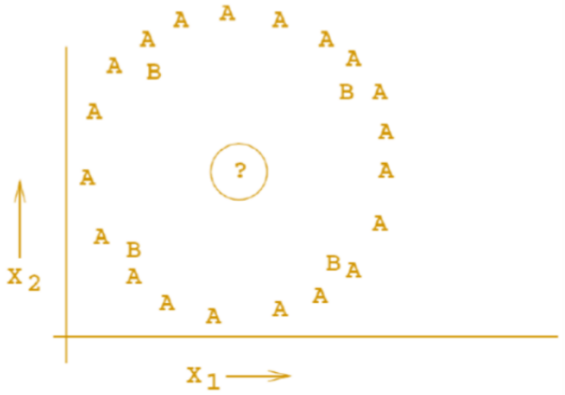
سوال ۵) دسته‌بندی‌های Naïve Bayes و Logistic Regression را با یک دیگر مقایسه کنید.(حداکثر یک صفحه)

سوال ۶) احتمال P(D|B=T) در شبکه بیزین زیر محاسبه کنید.

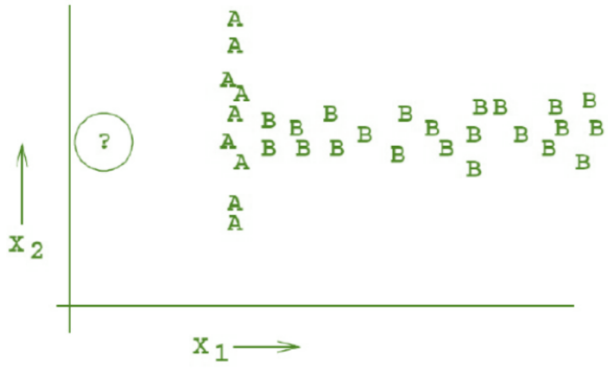


شکل ۲

سوال ۷) فرض کنید برای هر کدام از اشکال ۳ و ۴ یک دسته‌بند ساده گاوسی آموزش داده‌ایم. توضیح دهید که برچسب داده تست که با علامت سوال مشخص شده است چه خواهد بود.



شکل ۳

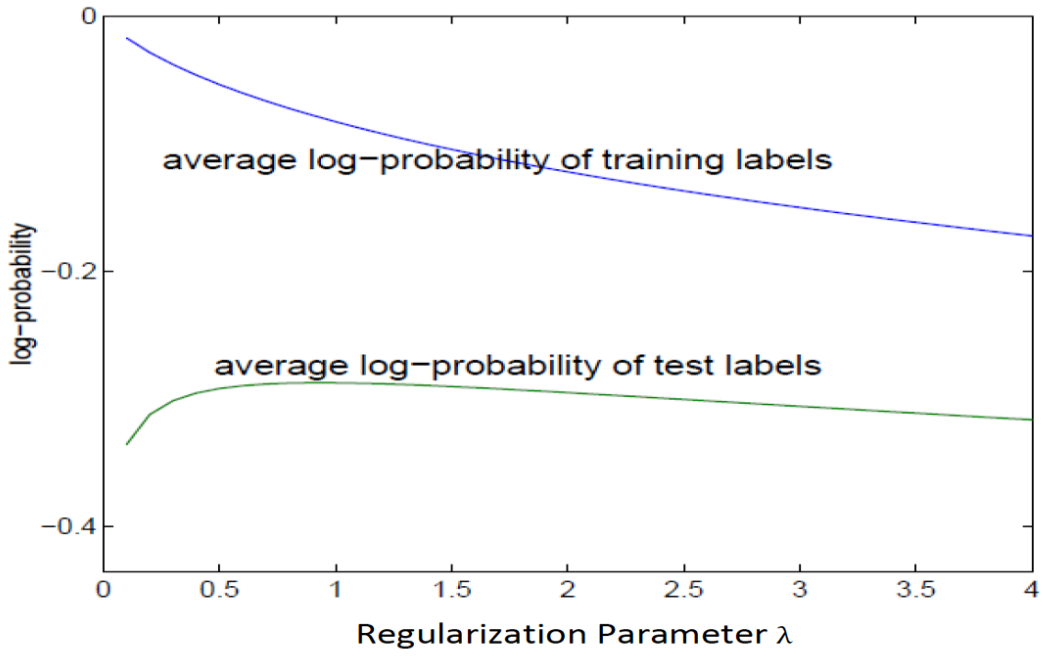


شکل ۴

سوال ۸) تابع هزینه زیر برای دسته‌بندی کننده Logistic Regression را در نظر بگیرید.

$$J(\theta) = -\left[\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \log h_\theta\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - h_\theta\left(x^{(i)}\right)\right)\right] + \frac{\lambda}{2m}\sum_{j=1}^{n} \theta_j^2$$

شکل ۵ میانگین لگاریتم احتمال برچسب داده‌های آموزشی و تست را بعد از آموزش دسته‌بند با ضریب تنظیم λ را نشان می‌دهد. توضیح دهید چرا با افزایش پارامتر λ، میانگین لگاریتم احتمال برچسب داده‌های تست کاهش می‌یابد؟



شکل ۵

- 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 K نزدیک‌ترین همسایه(KNN)

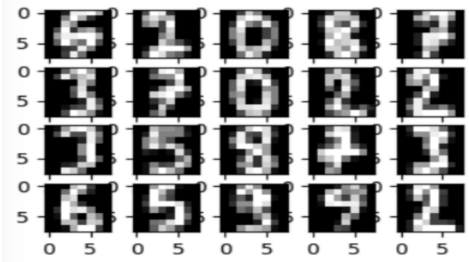
دیتاست این سوال mnist است که می‌توانید آن را از لینک زیر دانلود کنید:

http://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits

یا در کد پایتون می‌توانید آن را بصورت زیر استفاده کنید:

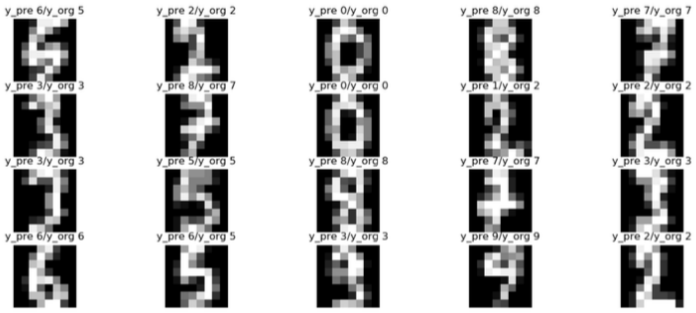
```
from sklearn import datasets

mnist = datasets.load_digits()
```

این دیتاست شامل ۱۷۹۷ تصویر از رقم‌های ۰ تا ۹ است که هر تصویر از ۶۴ پیکسل تشکیل شده که در شکل زیر این رقم‌ها را مشاهده می‌کنید.



شکل ۶

الف) ابتدا دیتاست را به سه مجموعه آموزش، تست و ارزیابی تقسیم کنید (مقادیر در نظر گرفتن اختیاری). سعی کنید در هر مجموعه تعداد یکسانی از اعضای کلاس‌های مختلف قرار داشته باشند. با استفاده از مجموعه‌ی آموزش و ارزیابی مقدارهای مناسبی برای k و تابع فاصله پیدا کنید. بعد از پیدا کردن بهترین مدل، خطای آموزش، ارزیابی و تست و همچنین ماتریس درهم‌ریختگی را گزارش کنید. سپس ۱۰۰ عضو از مجموعه‌ی تست به صورت تصادفی جدا کنید و با استفاده از KNN که آموزش داده‌اید کلاس هر رقم را پیش‌بینی کنید. ۱۰۰ رقم انتخاب شده، مقدار پیش‌بینی شده و مقدار واقعی را در شکلی مانند شکل۷ نمایش دهید.



شکل ۷

ب) قسمت الف را با استفاده از یک کتابخانه آماده تکرار کنید و نتایج را مقایسه کنید. پیشنهاد می‌شود از sklearn KNeighbhorsClassifier استفاده شود. استفاده از سایر کتابخانه‌ها مشکلی ندارد.

مسئله ۲) پیاده‌سازی دسته‌بند بیز ساده

الف) ابتدا مجموعه داده مربوطه را از لینک زیر دریافت کنید. پس از انجام پیش‌پردازش‌های لازم، دسته‌بند بیز ساده گاوسی را پیاده‌سازی کنید. به منظور گزارش دقت دسته‌بند از روش ۶-fold-cross-validation استفاده کنید (توجه نمایید که در صورت کوچک بودن احتمالات می‌توانید از لگاریتم احتمالات استفاده نمایید.)

https://archive.ics.uci.edu/ml/datasets/Wine

ب) قسمتی از داده‌ها را به عنوان داده تست در نظر گرفته و نمودار ROC را برای مدل آموزش داده رسم کرده و آن را تحلیل کنید.

مسئله ۳) پیاده‌سازی لاجستیک رگرسیون

مجموعه داده ارقام دست نویس MNIST که در مسئله ۱ از آن استفاده کردید، مجددا در این تمرین بکار گیرید. با استفاده از روش One-vs-All داده‌ها را دسته‌بندی کنید. برای این منظور می‌توانید از رگرسیون خطی یا غیر خطی استفاده کنید. در صورت استفاده از مدل غیر خطی درجه آن را به صورت دلخواه انتخاب کنید. در این تمرین نیازی به پیاده‌سازی رگرسیون لاجستیک

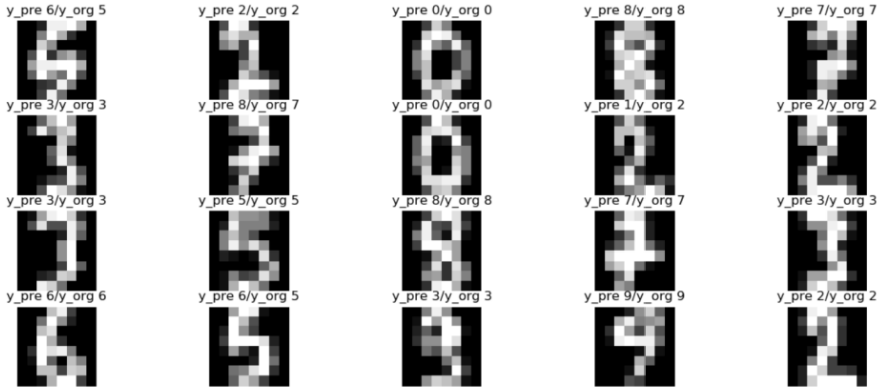
نیست. می‌توانید از کتابخانه آماده مانند sklearn استفاده کنید. اما توجه داشته باشید که قسمت One-vs-All باید پیاده‌سازی شود و از پارامتر multi-class استفاده نشود.

الف) پس از آموزش دسته‌بند، خطاهای مجموعه آموزش و تست و ماتریس درهم ریختگی را گزارش کنید.

ب) ۲۵ داده از مجموعه تست به صورت تصادفی انتخاب کرده و برای هر داده کلاس واقعی و کلاس پیش‌بینی شده توسط مدل آموزش داده شده در تصویری مانند تصویر ۸ گزارش کنید.

ج) عملکرد این روش را با روش K نزدیک‌ترین همسایه که در مسئله ۱ انجام دادید، مقایسه کنید.

د) یکی از مشکلاتی که می‌تواند در استفاده از روش One-vs-All رخ دهد، مشکل یادگیری نامتوازن است. این مسئله را به صورت مختصر توضیح دهید و بیان کنید برای حل این مشکل چه پیشنهادی دارید.



شکل ۸