دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر و فناوری اطلاعات

تمرین سری دوم درس یادگیری ماشین

پاییز ۱۳۹۶

بخش اول (KNN)

سؤال ۱

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

سؤال ۲

Bias و Variance الگوریتم KNN را با دلایل کافی مورد بررسی قرار دهید.

سؤال ۳

الگوریتم KNN، الگوریتمی پارامتری است، یا غیر پارامتری؟ صحت پاسخ ارایه شده را با دلایل کافی مورد بررسی قرار دهید.

سؤال ۴

مجموعه داده‌ی موجود در لینک زیر را دریافت کنید.

http://archive.ics.uci.edu/ml/datasets/seeds

با مجموعه داده دریافت شده به بخش‌های زیر پا سخ دهید. توابع مورد ا ستفاده را پیاده سازی کرده و از توابع و کتابخانه‌های آماده موجود استفاده نکنید.

أ) الگوریتم KNN با اندازه K های 1، 3، 5، 7 و 10 پیاده سازی کنید. ابتدا با استفاده از 10-fold cross validation دقت حاصل را اندازه‌گیری کرده و نمودار دقت حاصل را رسم کنید. نتایج به د ست آمده را تحلیل کنید. (توجه دا شته با شید که 10-fold cross validation باید پیاده سازی شود و امکان استفاده از توابع آماده وجود ندارد.)

ب) الگوریتم KNN را با پارامتر K = 5 و فاصله‌های اقلیدسی، فاصله منهتن، فاصله مینکوفسکی[1] با p = 4 و p = ½ و فاصله‌ی کسینوسی پیاده‌سازی کنید. هر کدام از فاصله‌های خواسته شده را در تابعی جداگانه پیاده سازی کنید. سپس با ا ستفاده از 10-fold cross validation دقت حاصل را اندازه گیری کنید. نتایج به دست آمده از این بخش را با توجه به ماهیت فاصله‌ها

[1] Minkowski distance

مورد تحلیل قرار دهید. (توجه دا شته با شید که 10-fold cross validation و فا صله‌های خواسته شده، باید پیاده سازی شود و امکان استفاده از توابع آماده وجود ندارد.)

سؤال ۵

مـقـالـه "An Improved KNN Algorithm Based on Minority Class Distribution for imbalanced Datasets" پیوست شده است.

أ) مزایا و معایب الگوریتم KNN که در مقاله پیو ست شده تو ضیح داده شده ا ست، ذکر کرده و مورد بررسی قرار دهید.

ب) برای حل معایب موجود چه راهکارهای پیشینی موجود است؟

ج) الگوریتم KNN و WDKNN در مقاله تو ضیح داده شده ا ست. الگوریتم WDKNN ن سبت به KNN دارای چه مزیتی است؟ الگوریتم WDKNN را پیاده‌سازی کنید.

د) روش پیشنهادی مقاله برای حل مشکل داده‌های نامتعادل (imbalanced) را توضیح داده و پیاده سازی کنید.

ه) نتایج حاصل از اجرای الگوریتم‌های پیاده‌سازی شده را بر روی ۷ مجموعه داده اول معرفی شده در مقاله، ثبت کرده و مورد بررسی قرار دهید.

بخش دوم (درخت تصمیم‌گیری)

سؤال ۱

یکی از مشکلات درخت تصمیم، احتمال بالای بیش‌برازش[2] این مدل است. در مورد این مشکل درخت تصمیم و دلایل آن توضیح دهید.

سؤال ۲

هرس کردن[3] درخت تصمیم به چه منظور صورت می‌گیرد؟

[2] Overfitting
[3] Pruning

سؤال ۳

جنگل تصادفی یکی از مدل‌هایی است که بر مبنای درخت تصمیم و با هدف برطرف کردن مشکلات آن ارائه شده‌است. توضیح کاملی از نحوه‌ی کارکرد این مدل در حداکثر یک صفحه ارائه دهید.

سؤال ۴

ابزار وکا[4] را از این لینک دریافت کنید. با توجه به مجموعه داده‌ی labor از مجموعه داده‌های نمونه‌ی وکا به موارد زیر پاسخ دهید.

أ) مجموعه داده‌ی مورد نظر را از تب preprocess بارگذاری کرده و از تب classify درخت تصمیم (J48) را با تنظیمات پیش‌فرض و با 10-fold cross validation آموزش داده و دقت[5]و ماتریس پریشانی[6] آن را گزارش کنید. سپس از روی ماتریس پریشانی مقادیر TP، FP، FN، TN، Precision، Recall و F1-Measure را بر حسب خانه‌های ماتریس پریشانی به دست آورید. درخت تصمیم ساخته شده را رسم کنید. داده‌ی 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

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؟

[4] Weka
[5] Accuracy
[6] Confusion Matrix

ج) در این قسمت از جنگل تصادفی استفاده کرده و تمام موارد خواسته شده در بخش (أ) (به جز دسته‌بندی داده‌ی ارائه شده) را گزارش کنید. تعداد درخت‌ها و ویژگی‌های در نظر گرفته شده را نیز گزارش کنید.

د) کدام یک از مدل‌های قسمت‌های قبل عملکرد بهتری داشته‌اند؟ دلیل این امر چه می‌تواند باشد؟

توضیحات تمرین:

۱- شما باید سورس کد خود به همراه گزارش (پاسخ سؤال‌ها و نتایج و تحلیل پیاده‌سازی‌هایی که خواسته‌شده‌اند) را در قالب یک فایل Zip با نام فایل $xxxxx\_hw2$ که $xxxxx$ شماره دانشجویی شما است، تا تاریخ ۲۶ آبان در سایت درس بارگذاری کنید.

۲- پیاده‌سازی با متلب یا پایتون باید انجام شود.

۳- مجاز به استفاده از هیچ کتابخانه آماده‌ای نیستید.

۴- در صورت هرگونه سؤال یا ابهام به ceit17@gmail.com ایمیل بزنید.