یادگیری ماشین

پروژه پایانی

نگارش

حامد زارعی

نیمسال اول ۹۸-۹۷

۱. تعریف مساله

در این پروژه برای کار کردن و یاد گرفتن الگوریتم‌هایی مانند نزدیک ترین همسایه، نزدیک‌ترین همسایه به صوت وزن‌دار، شبکه RBF و همچنین درخت تصمیم است. برای این کار یکسری داده از داده-های معروف مورد کاربرد در مقاله‌ها استفاده شده است.

در این پروژه روند کار به این صورت است که برای تمام داده‌ها ابتدا با استفاده از PCA ابعاد مسئله را کاهش می‌دهیم سپس با قسمت بندی داده با استفاده از مفهوم 5-fold cross validation داده‌های آموزش و ارزیابی درست می‌کنیم و هرکدام از الگوریتم‌های مطرح شده را بر روی آن تست می‌کنیم. در نهایت معیارهایی همچون میانگین دقت و یا حساسیت و دیگر معیارها را بدست می‌آوریم.

۲. حل مساله

برای این پروژه برای تمام ترکیب داده‌ها و روش‌ها یک کد در نظر گرفته شده است که به ترتیب قدم‌های زیر را برای آن طی شده است:

۱) کاهش ابعاد
۲) محاسبه مقدار هر بخش از fold ها
۳) ایجاد loop برای محاسبه هر بخش و بدست آوردن مقدارهای کلی مورد نیاز
٤) محاسبه معیارهای ارزیابی مورد نیاز

۲,۱. نزدیک‌ترین همسایه

در این روش با ورود هر نمونه جدید براساس داده‌های موجود و فاصله‌ای که نسبت به هریک از نمونه‌ها دارد یک برچسب داده می‌شود. به این صورت که فاصله نمونه جدید با تمام نمونه‌های موجود بررسی شده و براساس نوع و تعداد همسایگی (k) نمونه‌ای که بیشتر تکرار در آن مجموعه را داشته باشد انتخاب می‌شود.

۲,۲. نزدیک‌ترین همسایه وزن‌دار

این روش مانند روش قبل است با این تفاوت که در تعداد نمونه نزدیک‌تر انتخاب شده فقط تعداد بیشتر اهمیت ندارد و فاصله از نمونه جدید به عنوان وزنی اعمال می‌شود. به این صورت که در صورتی که مثلا در k=3 اگر از یک کلاس دو بار تکرار ولی با فاصله‌های زیاد باشد ولی از یک کلاس خیلی نزدیک به نمونه موجود باشد، آن کلاس به عنوان کلاس نمونه جدید ورودی انتخاب خواهد شد.

۲,۳. شبکه RBF

در این شبکه سعی بر آن است که داده‌هایی که به صورت خطی جدا پذیر نیستند با تبدیل‌هایی به فضاهای جدیدی ببریم تا بتوان آن‌ها را از هم به صورت خطی جدا کرد.

۲,۴. درخت تصمیم

خروجی در این روش یک درخت خواهد بود. در ریشه این درخت پراهمیت ترین ویژگی قرار خواهد گرفت. پراهمیت ترین یعنی بهره اطلاعاتی آن از بقیه بیشتر باشد.

۳. توابع آزمون – مجموعه داده

توابع آزمون موارد زیر خواهد بود:

- دقت: تعداد درست تشخیص داده‌ها به تعداد کل داده
- Sensitivity
- Specificity

$$\mathbf{S}ensitivity = \frac{TP}{TP + FN} = \frac{TP}{P}$$

$$\mathbf{S}pecificity = \frac{TN}{TN + FP} = \frac{TN}{N}$$

- F1

$$F_1 = 2 \cdot \frac{\text{Precison} \times \text{Recall}}{\text{Precison} + \text{Recall}}$$

$$F_1 = \frac{2TP}{P + P'} = \frac{2TP}{2TP + FP + FN}$$

- MCC

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

$$= \frac{TP \times TN - FP \times FN}{\sqrt{PNP'N'}}$$

مجموعه داده‌ها عبارتند از:

- Glass: ۱۰ ویژگی، ۲۱٤ نمونه در ۷ کلاس، دسته بندی شده اند.
- Diabetic Indians Pima: ۸ ویژگی، ۷٦۸ نمونه در ۲ کلاس، دسته بندی شده اند.
- Statlog Heart: ۱۳ ویژگی، ۲۷۰ نمونه در ۲ کلاس، دسته بندی شده اند.
- Breast Cancer: ۳۰ ویژگی، ٥٦۹ نمونه در ۲ کلاس، دسته بندی شده اند.

۴. شرح عملکرد برنامه

برای پیاده سازی این پروژه از متلب ۲۰۱٤ استفاده شده است در یک ماشین با ۸ گیگ رم.

حل هر بخش از مسئله به صورت جداگانه در فایلی آمده است.

۴,۱. مثالی از معرفی یک زیر برنامه – رسم نمودار

```
function labels = rbf( train, test )
```

. تابعی برای پیاده سازی شبکه rbf به صورت دستی و با استفاده از الگوریتم اسلایدها به این صورت که مقادیر آموزش و تست را گرفته و خروجی کلاس‌های مرتبط به دسته تست خواهد بود.

۵. شبیه سازی ها و نتایج

جدول ۱ نتایج بررسی و تست بر روی داده‌های Glass آمده است.

جدول ۱: پارامترهای استفاده شده در جستجوی تپه نوردی

Method	***Mean***	***Best***	***Worst***	***STD***	***Time***
3NN	0.8952	0.9523	0.8333	0.0432	0.8759
5NN	0.8714	0.9524	0.8333	0.0494	0.6099
DW-3NN	0.9190	0.9762	0.9048	0.0319	0.6459
DW-5NN	0.9000	0.9762	0.8571	0.0458	0.6022
RBF-Code	0.2048	0.2857	0.1190	0.0782	0.2157
RBF-Toolbox	0.8190	0.8571	0.7857	0.0319	51.2844
Tree	0.6299				0.05

```
=== Classifier model (full training set) ===

Decision Stump

Classifications

a1 <= 0.7629107980000001 : 1.6748466257668713
a1 > 0.7629107980000001 : 6.313725490196078
a1 is missing : 2.7803738317757007


Time taken to build model: 0.05 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.9412
Mean absolute error                      0.6299
Root mean squared error                  0.7089
Relative absolute error                 36.6099 %
Root relative squared error             33.7188 %
Total Number of Instances              214
```

جدول داده‌های pima

Method	***Mean***	***Best***	***Worst***	***STD***	***F1***	***Sensitivity***	***Specificity***	***MCC***	***Time***
3NN	0.7294	0.6667	0.7712	0.0407	08195	0.8485	0.5926	0.4569	0.7233
5NN	0.7399	0.7582	0.6601	0.0526	0.8406	0.8788	0.6111	0.5138	0.6099
DW-3NN	0.7229	0.9762	0.6471	0.0404	0.8155	0.8485	0.5741	0.4404	0.6675
DW-5NN	0.7320	0.9762	0.8571	0.0458	0.8293	0.8586	0.6111	0.4886	0.5754
RBF-Code	0.6471	0.7320	0.5882	0.0533	0.7823	0.9798	0.0370	0.0504	0.2185
RBF-Toolbox	0.7307	0.7582	0.6667	0.0365	0.8203	0.8990	0.4630	0.4118	41.1404
Tree	0.6299								0.05

خروجی درخت‌ها عکس گرفته شده است.

جدول داده‌های Heart

Method	***Mean***	***Best***	***Worst***	***STD***	***F1***	***Sensitivity***	***Specificity***	***MCC***	***Time***
3NN	0.7778	0.8333	0.7407	0.0346	0.7797	0.7667	0.7500	0.5149	0.5505
5NN	0.7399	0.7582	0.6601	0.0526	0.8406	0.8788	0.6111	0.5138	0.6099
DW-3NN	0.7229	0.9762	0.6471	0.0404	0.8155	0.8485	0.5741	0.4404	0.6675
DW-5NN	0.7320	0.9762	0.8571	0.0458	0.8293	0.8586	0.6111	0.4886	0.5754
RBF-Code	0.6471	0.7320	0.5882	0.0533	0.7823	0.9798	0.0370	0.0504	0.2185
RBF-Toolbox	0.7307	0.7582	0.6667	0.0365	0.8203	0.8990	0.4630	0.4118	41.1404
Tree	0.6299								0.05

جدول داده‌های Cancer

Method	***Mean***	***Best***	***Worst***	***STD***	***F1***	***Sensitivity***	***Specificity***	***MCC***	***Time***
3NN	0.9575	0.9823	0.9381	0.0192	0.9600	0.9474	0.9867	09403	0.6634
5NN	0.7399	0.7582	0.6601	0.0526	0.8406	0.8788	0.6111	0.5138	0.6099
DW-3NN	0.7229	0.9762	0.6471	0.0404	0.8155	0.8485	0.5741	0.4404	0.6675
DW-5NN	0.7320	0.9762	0.8571	0.0458	0.8293	0.8586	0.6111	0.4886	0.5754
RBF-Code	0.6471	0.7320	0.5882	0.0533	0.7823	0.9798	0.0370	0.0504	0.2185
RBF-Toolbox	0.7307	0.7582	0.6667	0.0365	0.8203	0.8990	0.4630	0.4118	41.1404
Tree	0.6299								0.05