بنام خدا

شناسایی الگو دکتر ابوالقاسمی

توجه: در این تمرین از داده‌های tiny MNIST استفاده خواهیم کرد.

سوال ۱) برای داده‌های داده شده، ابتدا یک طبقه‌بند بهینه‌ی بیز با روش پنجره‌ی پارزن برای تخمین pdf با حالت‌های زیر طراحی کنید:

۱. دو پنجره‌ی مستطیلی و گوسی را بررسی کنید.

۲. تاثیر اندازه‌ی پنجره را بررسی کنید. (۳ حالت مختلف را در این قسمت بررسی کنید.)

نرخ طبقه‌بندی صحیح (CCR) و احتمال خطا را گزارش کنید.

سوال ۲) سوال قبل را با استفاده از روش K نزدیکترین همسایه برای تخمین pdf دوباره تکرار کنید. برای سه مقدار مختلف k الگوریتم را تکرار کرده و نتایج را گزارش کنید. (مقادیر را طوری انتخاب کنید که نتایج تفاوت زیادی باهم داشته باشند و استدلال کنید.) حد بالا و پایین خطا را برای این تخمین بدست آورید. (در صورت طولانی شدن الگوریتم، تعدادی از داده‌ها را انتخاب کرده و الگوریتم را بر روی آن‌ها اجرا کنید. در اینصورت حتما در گزارش خود تعداد داده‌ها را نیز ذکر کنید.)

سوال ۳) با افزایش مقادیر k و n برای هر دو تخمین بالا، همگرا شدن نتایج به همدیگر را در دو روش مشاهده کنید.

سوال ۴) برای سرعت بخشیدن به الگوریتم k نزدیکترین همسایه، آن را یکبار دیگر با k-d tree پیاده‌سازی کنید. سپس درجه‌ی محاسباتی را با توجه به تعداد داده‌ها، برای حالت بدون k-d tree و با k-d

tree به صورت تقریبی بدست آورید.

سوال ۵) در این سوال از طبقه‌بند بیز استفاده کنید. (به عنوان متری برای مقایسه‌ی خطای هر مرحله از مجموع مربعات خطا استفاده کنید. همچنین برای طبقه‌بند بیز می‌توانید از پکیج sklearn استفاده کنید.)

۱. الگوریتم Naive search را برای داده‌های داده شده پیاده‌سازی کنید.

۲. الگوریتم Sequential Forward Selection را برای داده‌های داده شده پیاده‌سازی کنید.

۳. الگوریتم Sequential Backward Elimination را برای داده‌های داده شده پیاده‌سازی کنید.

۴. مزیت Backward Elimination را نسبت به Forward Selection به صورت کامل شرح دهید.

۵. برای هر سه الگوریتم بالا، اندیس بهترین ویژگی‌ها و همچنین نمودار خطا برحسب تعداد ویژگی‌ها را رسم کنید. سپس سرعت و دقت هر سه الگوریتم را مقایسه کنید.

۶. کدام یک از سه الگوریتم ذکر شده به PCA نزدیک است. همچنین بحث کنید که با اعمال چه تغییراتی، الگوریتم مذکور به PCA تبدیل می‌شود.

سوال ۶) با توجه به اینکه در قسمت قبل، دو الگریتم Forward Selection و Backward Elimination را پیاده‌سازی کرده‌اید، حال به کمک این دو الگوریتم Sequential Floating Forward Selection را نیز پیاده‌سازی کرده و نتایج بدست آمده را با قسمت قبلی مقایسه کرده و استدلال کنید.

سوال ۷) الگوریتم PCA را برای داده‌های داده شده پیاده‌سازی کنید. ابتدا مقادیر ویژه را به ترتیب کاهشی رسم کرده و سپس با توجه به نمودار کاهش خطا برحسب تعداد ویژگی‌های انتخاب شده، بهترین تعداد ویژگی را انتخاب کنید.

با توجه به این که طبقه‌بند بیز را با استفاده از دو تخمینگر pdf پنجره‌ی پارزن و k نزدیکترین همسایه بدست آوریده‌اید، بهترین پارامترهای سوال ۱ و ۲ را انتخاب کرده و بار دیگر این قسمت‌ها را پس از اعمال PCA تکرار کنید و نتایج (CCR) را مقایسه کنید.

سوال ۸) روش Information theoretic Feature Selection را توضیح دهید. یک روش برای انجام این کار استفاده از Mutual Information بین ویژگی‌های مختلف است. روش اجرای این حالت را به صورت الگوریتمی بیان کنید. همچنین مزایا و معایب این روش را هم توضیح دهید. [۱]

سوال ۹) فرض کنید که پنجره‌ی پارزن بصورت $\phi(x) = e^{-x}$ برای $x > 0$ و صفر برای $x \leq 0$ تعریف شده است. برای $p(x) \sim Uniform(0, a)$ نشان دهید که میانگین پنجره‌ی پارزن بصورت زیر است:

$$p(x) = \begin{cases} 0 & ; x < 0 \\ \frac{1}{a}(1 - e^{\frac{-x}{h_n}}) & ; 0 \leq x \leq a \\ \frac{1}{a}(e^{\frac{a}{h_n}} - 1)e^{\frac{-x}{h_n}} & ; a \leq x \end{cases}$$

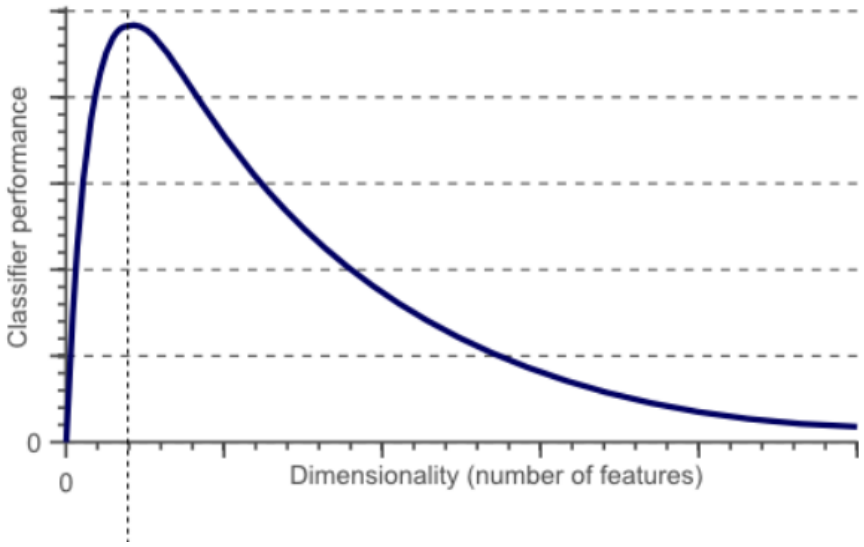
سوال ۱۰) یکی از کاربردهای PCA استفاده از آن در eigenfaces است. در این سوال قصد داریم تا از این کاربرد استفاده کنیم. ابتدا دلیل استفاده از PCA برای این کار را توضیح دهید. سپس بر روی داده‌های Labeled Faces in the Wild (LFW) و با استفاده از یک طبقه‌بند بیز، این الگوریتم را پیاده‌سازی کنید. برای دانلود و بارگذاری داده‌ها می‌توانید از دستور زیر استفاده کنید.

[۱] http://www.jmlr.org/papers/v13/brown12a.html

```
from sklearn.datasets import fetch_lfw_people
dataset = fetch_lfw_people(min_faces_per_person=100)
_, h, w = dataset.images.shape
data = dataset.data
labels = dataset.target
target_names = dataset.target_names
```

دقت کنید که در این سوال هدف رسیدن به دقت بالا نیست. بلکه هدف سرعت بخشیدن به الگوریتم با استفاده از الگوریتم PCA است. در این سوال هم برای پیاده‌سازی الگوریتم PCA و هم برای پیاده‌سازی طبقه‌بند بیز می‌توانید از پکیج sklearn استفاده کنید.

سوال ۱۱) (امتیازی) ابتدا داده‌های Fashion-MNIST را بارگذاری کنید. ۱۰۰۰۰ داده را بصورت تصادفی انتخاب کرده و سعی کنید با اعمال PCA و کاهش ابعاد داده‌ها به اعداد مختلف منحنی زیر را که مربوط به Curse of Dimensionality است، بدست آورید. از طبقه‌بند بهینه‌ی بیز استفاده کنید. (توجه داشته باشید که در این سوال می‌توانید از پکیج sklearn استفاده کنید.)



لطفا به نکات زیر توجه داشته باشید:

۱. کدهای مربوط به هر تمرین را به صورت py. یا m. ارسال کنید و از ارسال فایل با فرمت ipynb. خودداری کنید.

۲. تمامی نکات لازم و فرض‌های خود در برنامه نویسی را در گزارش خود ذکر کنید.

۳. داده‌های در نظر گرفته شده برای این سری تمرین tiny MNIST است. فقط در سوال ۱۰ از داده‌های دیگری استفاده خواهیم کرد.

۴. در این این تمرین مجاز به استفاده از پکیج‌های آماده نیستید. مگر در مواردی که در متن سوال به صراحت استفاده از پکیج مجاز اعلام شده است.

۵. فایل ارسالی شامل گزارش با فرمت pdf و پوشه‌ی کدها خواهد بود.

۶. نام فایل ارسالی حتما شامل نام، نام‌خانوادگی و شماره‌ی دانشجویی شما باشد.

۷. در صورت داشتن هر گونه سوالی با ایمیل زیر در ارتباط باشید:

pouya.narimani@ut.ac.ir

موفق و پیروز باشید