دانشکده فنی دانشگاه تهران

دانشکده برق و کامپیوتر

تمرین ۴ سیستم‌های هوشمند

طراحان:

امین فدایی‌نژاد

میرحامد جعفرزاده

رایانامه

aminfadaei116@gmail.com

jafarzadeh.mirhamed@gmail.com

نیم سال اول ۱۳۹۹-۱۴۰۰

دانشجویان عزیز، قبل از پاسخ‌گوئی به سوالات به نکات زیر توجه کنید:

۱. شما باید کدها و گزارش خود را با الگو IS_HW4_StudentNumber.zip در محل تعیین شده آپلود کنید.

۲. گزارش کار شما نیز از معیارهای ارزیابی خواهد بود، در نتیجه زمان کافی برای تکمیل آن اختصاص دهید.

۳. شما میتوانید سوالات خود را از طریق ایمیل طراحان تمرین بپرسید.

۱. روش خوشه کردن k-means یک خوشه بندی بهینه محور [۱] می‌باشد که در آن تابع هزینه می‌بایست کمینه گردد. هدف آن تقسیم بندی تعداد n داده به k خوشه می‌باشد که در هر کدام از این دسته‌ها تعدادی داده متشابه وجود دارد. شما باید با مشاهده داده‌های $(x_1, x_2, ..., x_m)$ آن‌ها را به خوشه‌های $S = \{S_1, S_2, ..., S_k\}$ تقسیم کنید به نحوی که تابع WCSS [۲] کمینه گردد.

اصولا کمینه کردن تابع WCSS به معنا پیدا کردن دسته خوشه S می‌باشد به نحوی که تابع زیر کمینه شود:

$$CostFunction = \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||_2^2$$

که در این رابطه μ_i برابر میانگین داده‌های موجود در خوشه S_i می‌باشد.

روش پیاده‌سازی الگوریتم:

a. تعیین مقدار اول $k, \mu_1, \mu_2, ..., \mu_k$

b. تخصیص داده‌ها به به خوشه‌های با نزدیک ترین μ_i

c. محاسبه مجدد μ_iها

d. تا جایی که مقدار μ_iها ثابت شود

e. بازگرداندن مقدار μ_iها

f. پایان

- الگوریتم فوق را from scratch پیاده‌سازی کنید (نمی‌توانید از کتابخانه‌های آماده استفاده کنید).
- با استفاده از کتابخانه‌های آماده، قسمت قبل را پیاده‌سازی کنید. (scikit-learn)
- برای دو قسمت قبل Confusion Matrix را رسم کنید و نتیجه را با هم مقایسه کنید.

[۱] Optimization-based Clustering
[۲] Within-cluster sum of squares

۲. فرض کنید می‌خواهیم ۵ نقطه با مختصات مشخص را به شکل سلسله مراتبی [۱] خوشه‌بندی کنیم. برای این کار از دو روش Agglomerative و Divisive استفاده می‌کنیم. در هر یک از این دو روش معیارهای فاصله جهت انجام محاسبات مشخص شده است. توجه کنید که برای هر روش و معیار فاصله، Dendogram را نیز نمایش دهید.

- روش Agglomerative :
  - معیار Complete-Link : $cd(X,Y) = \max_{x \in X, y \in Y} d(x,y)$
  - معیار Centroid : $cd(X,Y) = d(avg(X), avg(Y))$
- روش Divisive :
  - معیار Single-Link : $cd(X,Y) = \min_{x \in X, y \in Y} d(x,y)$

$$p_1 = \begin{pmatrix} 12 \\ 9 \\ 7 \end{pmatrix} \quad \& \quad p_2 = \begin{pmatrix} -2 \\ 4 \\ 4 \end{pmatrix} \quad \& \quad p_3 = \begin{pmatrix} 15 \\ 0 \\ 1 \end{pmatrix} \quad \& \quad p_4 = \begin{pmatrix} 3 \\ -1 \\ 2 \end{pmatrix} \quad \& \quad p_5 = \begin{pmatrix} 11 \\ 4 \\ 9 \end{pmatrix}$$

- برای این سوال به کد زدن نیازی نیست.

[۱] Hierarchical

۳. در این سوال، می‌خواهیم با استفاده از روش طبقه‌بند بیزی [۱] (و سپس با Risk Minimization) یک طبقه‌بند باینری را بر اساس توزیع احتمال شرطی برچسب [۲] نسبت به داده‌ی مشاهده شده‌ی x بیابیم.
توابع توزیع احتمال به شکل زیر است:

$$p(w_1|x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{(x)^2}{2})$$

$$p(w_2|x) = \frac{1}{2\sqrt{2\pi}} \exp(-\frac{(x-4)^2}{8})$$

- با استفاده از توزیع‌های داده شده، مقدار x_0 (یعنی نقطه‌ی تعیین‌کننده‌ی نواحی تصمیم‌گیری) را بدست آورید.
- با در نظر گرفتن ماتریس λ به عنوان Conditional Risk ، روش Risk Minimization را به طبقه‌بند قسمت قبل اضافه کرده و مقدار جدید x_0 را بدست آورید. دلیل تغییر مقدار x_0 را به شکل مفهومی توجیه کنید.
راهنمایی: به مفهوم ماتریس λ که در درس به آن اشاره شده است، توجه کنید.
- برای این سوال به کد زدن نیازی نیست.

$$\lambda = \begin{pmatrix} 0 & 2 \\ 1 & 0 \end{pmatrix}$$

[۱] Bayesian Classifier
[۲] Label

۴. در این سوال قصد داریم با استفاده از Bayes Optimal classifier توزیع داده‌های موجود را با با توزیع گوسی [۱] تخمین بزنیم. همچنین می‌دانیم که توزیع گوسی در فضا چند بعدی به صورت زیر تعریف می‌شود:

$$p(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_i|}\exp(-\frac{1}{2}(x-\mu_i)^T\Sigma^{-1}(x-\mu_i))$$

که در رابطه فوق μ_i میانگین داده‌های برچسب iام می‌باشد. Σ_i ماتریس کوواریانس [۲] داده‌های کلاس iام و n نیز بعد داده‌های ما می‌باشد. می‌توان دو پارامتر مذکور را از این طریق به درست آورد:

$$\mu_i = \frac{1}{Q_i}\sum_{x_j \in C_i} x_i \qquad \Sigma_i = \frac{1}{Q_i - 1}\sum_{x_j \in C_i}(x_j - \mu_i)(x_j - \mu_i)^T$$

پارامتر دانش اولیه یا همان Prior Knowledge را می‌توان از طریق فرمول زیر به دست آورد:

$$P_i = \frac{Q_i}{\sum_{j=1}^{numclass} Q_j}$$

- با استفاده از طبقه‌بند فوق داده‌های Test را طبقه‌بندی کنید. (بدون در نظر گرفتن پارامتر دانش اولیه [۳])
- قسمت قبل را با داشتن پارامتر دانش اولیه تکراری کنید.
- برای دو قسمت قبل Confusion Matrix را رسم کنید و نتیجه را با هم مقایسه کنید

[۱] Gaussian
[۲] Covariance Matrix
[۳] Prior Knowledge

۵. در این سوال قصد داریم با استفاده از روش Naive Bayes ، میزان درآمد اشخاصی را که در مجموعه داده [۱] ی income_Q5.csv قرار دارند، تخمین بزنیم. در اصل این کار را با دسته‌بندی اشخاص بر حسب ۲ مقدار ممکن برای درآمد آن‌ها انجام می‌دهیم.

- با استفاده از منطق ۸۰-۲۰ داده‌ها را به دو مجموعه‌ی آموزش [۲] و آزمایش [۳] تقسیم کنید و دقت طبقه‌بند را بدست آورید.
- برای قسمت قبل Confusion Matrix را رسم کنید.
- دقت کنید که طبقه‌بندی بر اساس میزان درآمد [۴] صورت می‌گیرد.
- حتما از روش Laplace Smoothing استفاده کنید. علت اهمیت این روش را به اختصار توضیح دهید.

موفق باشید

[۱] Dataset
[۲] Train
[۳] Test
[۴] Income