سلام بر تمام دانشجویان عزیز، چند نکته مهم:

1. حجم گزارش به هیچ عنوان معیار نمره‌دهی نیست، در حد نیاز توضیح دهید.
2. نکته‌ی مهم در گزارش نویسی روشن بودن پاسخها می‌باشد، اگر فرضی برای حل سوال استفاده می‌کنید حتما آن را ذکر کنید، اگر جواب نهایی عددی است به صورت واضح آن را بیان کنید.
3. برای سوالات شبیه سازی، فقط از دیتاست داده شده استفاده از کنید. شکل ها به طور واضح و در فرمت درست گزارش شوند.
4. حداکثر تا نمره ۱۱۰ (۱۰ نمره امتیازی) لحاظ خواهد شد.
5. هرگونه شباهت در گزارش و کد مربوط به شبیه سازی، به منزله تقلب می باشد و کل نمره تمرین صفر می‌شود.
6. در صورت داشتن سوال، برای بخش اول تمرین از طریق ایمیل mesbahamirhossein@gmail.com و برای بخش دوم از طریق ایمیل Rezatalakoob@yahoo.com، سوال خود را مطرح کنید.

تمرین سوم درس یادگیری ماشین

پاییز ۱۴۰۰

بخش اول) Dimension Reduction

1) سوال اول (۲۵ نمره)

الگوریتم backward elimination و Forward Selection را بر روی دیتاست TinyMNIST پیاده‌سازی کنید. برای طبقه بندی می‌توانید از Naïve Bayes optimal classifier استفاده کنید. (برای الگوریتم backward elimination و Forward Selection مجاز به استفاده از کتابخانه یا پکیج آماده نیستید ولی برای طبقه بند می‌توانید از پکیج‌های آماده استفاده کنید.)

مقدار CCR را بر حسب تعداد ویژگی‌های انتخاب شده در یک نمودار رسم کنید. همچنین تعداد بهینه ویژگی‌ها را برای بهترین عملکرد طبقه‌بند بیان نمایید.

2) سوال دوم (۱۵ نمره)

در این سوال میخواهیم به بررسی روش Pca بپردازیم. در دیتاست TinyMNIST لیبل کلاس‌ها را در نظر گرفته و مقادیر و بردارهای ویژه کواریانس را حساب کنید.

مقادیر ویژه ماتریس کواریانس را بر حسب شماره ویژگی رسم کنید. همچنین همانند بخش قبل تعداد بهینه ویژگی‌ها را بر اساس بخش پیشین بیان نمایید. سپس با انتقال داده‌ها به زیر فضای جدید که فقط شامل ویژگی‌های بهینه هستند، طبقه بند naïve Bayes optimal classifier با تخمین پارامتر گوسی را پیاده‌سازی کنید و مقدار CCR را گزارش کنید.

نتیجه خود را با مقدار CCR به دست آمده برای ویژگی‌های بهینه در قسمت قبل مقایسه کنید.

3) سوال سوم (۱۵ نمره)

عبارت

$$J = \frac{1}{n_1 n_2} \sum_{y_i \in Y_1} \sum_{y_j \in Y_2} (y_i - y_j)^2$$

پراکندگی کلی درون گروهی (within group scatter) را اندازه می‌گیرد. نشان دهیدکه این عبارت را میتوان به صورت زیر هم نوشت:

$$J = (m_1 - m_2)^2 + \frac{1}{n_1} {s_1}^2 + \frac{1}{n_2} {s_2}^2$$

4) سوال چهارم (۲۰ نمره)

در مساله‌ی طبقه‌بندی c کلاسه، ماتریس پراکندگی درون کلاسی و بین کلاسی به ترتیب به صورت زیر تعریف میشود:

$$S_w = \sum_{k=1}^{c} \sum_{x^q \in w_i} (x^q - \mu_k)(x^q - \mu_k)^T$$

$$S_B = \sum_{k=1}^{c} N_k (\mu_k - \mu)(\mu_k - \mu)^T$$

الف) نشان دهید $rank(S_B) \leq C - 1$ و در چه شرایطی $rank(S_B) = C - 1$ ؟

ب) درباره حداکثر تعداد مقادیر ویژه ناصفر ماتریس جداپذیری ${S_W}^{-1} S_B$ بحث نمایید.

ج) نشان دهید: $S_T = S_W + S_B$

بخش دوم) Linear discriminant functions & SVM

5) سوال پنجم (۱۰ نمره)

با توجه به تابع هزینه ی زیر برای مسئله ی soft margin Svm نشان دهید که مسئله ی class-separable بوده و با تشکیل ترم لاگرانژین و بررسی شرایط بهینگی ، جواب بدست آمده برای وزن ها را با جواب بهینه ی مسئله ی اصلی مطرح شده در کلاس مقایسه نمایید.

$$J(w) = \frac{1}{2} \| w \|_2^2 + \frac{c}{2} \sum_{i=1}^{N} \xi_i^2$$

$$S.t : y_i (w^T x_i + b) \geq 1 - \xi_i$$

---

6) سوال ششم (۱۵ نمره)

الف) تفاوت بین رویکردهای generative و discriminative را برای مسائل طبقه بندی توضیح دهید.

ب) مزایا و معایب روش های one-vs-one و one-vs-rest و linear machine را نسبت به همدیگر بیان کنید.

ج) توضیح دهید که روش dual problem برای حل مسئله ی بهینه سازی Svm چه مزیتی نسبت به روش مستقیم مسئله اولیه (primal problem) دارد.

7) سوال هفتم (۲۰ نمره)

الف) نشان دهید که برای کرنل $K(x_i, x_j) = \exp(-\frac{1}{2} \| x_i - x_j \|^2)$ به ازای هر دو ورودی دلخواه در فضای feature space خواهیم داشت(۵ نمره) :

$$| \phi(x_i) - \phi(x_j) \| \leq 2$$

ب) اگر کرنل های معتبر $k_1(x, y), k_2(x, y)$ را داشته باشیم، اعتبار کرنل های زیر را به کمک تئوری mercer بررسی نمایید. (۱۵ نمره)

1) $K(x, y) = f(x) K_1(x, y) f(y)$ برای هر f دلخواه

2) $K(x, y) = x^T A y$ که در آن A یک ماتریس معین مثبت می باشد.

3) $K(x, y) = (x^T y + 1)^p \quad p > 0$

4) $K(x, y) = x - x^T y$

5) $K(x, y) = \exp(K_1(x, y))$

8) سوال هشتم (۳۰ نمره)

در این سوال به اعمال طبقه بندی به کمک Support vector machines بر روی مجموعه داده Iris کار خواهیم پرداخت.(در این سوال امکان استفاده از sklearn برای بخش های مختلف را دارید.)

```
from sklearn import datasets
iris= datasets.load_iris()
```

ابتدا بر حسب دو ویژگی Petal Width و Petal Length داده های هر کلاس را مشخص نمایید و نمودار آن ها رسم کنید. حال با اعمال Svm در هر مرحله و تحت شرایط ذکر شده، نتایج از قبیل confusion matrix و دقت و f1-score را بیان نمایید.

الف) در مورد کرنل های rbf و linear و polynomial تحقیق کنید و بیان کنید هر کدام برای طبقه بندی چه مجموعه داده ای مناسب هستند. سپس این طبقه بند ها را اعمال کرده و نتایج را ذکر کنید.

ب) در مورد هایپرپارامترهای Gamma و Regularization تحقیق کنیدو هر کدام از این هایپرپارامترها را سه مرتبه تغییر دهید و طبقه بند را اعمل کنید. تاثیر هر پارامتر را تحلیل نمایید.

ج) به کمک gird search بهترین پارامتر ها را محاسبه نماییدو برای کرنل های ذکر شده طبقه بند را اعمال نمایید و نتایج را تحلیل کنید.

د) در این بخش رویکردهای مختلف برای مسئله چندکلاسه را برای سه کرنل linear و rbf و polynomial(d=3) را استفاده کنید و نتایج را تحلیل نمایید