شناسایی الگو
استاد: دکتر ابوالقاسمی

تمرین سری اول — طراح: پویا نریمانی

۱) (کدهای این سوال باید ضمیمه شود.)در این سوال قصد داریم تا ممان‌های داده‌های تولید شده از یک توزیع نرمال را بدست آوریم. سپس با استفاده از دو نوع طبقه‌بند مختلف تعدادی داده‌ی تصادفی را طبقه‌بندی کنیم.

۱. ابتدا دو دسته داده‌ی مختلف با استفاده از دو توزیع نرمال، اولی با میانگین ۱- و واریانس ۲ و دومی با میانگین ۴ و واریانس ۲.۲۵ تولید کنید.(تعداد داده‌های هر دسته را ۱۰۰ در نظر بگیرید.)

۲. دو توزیع تولید شده بر حسب مقادیر آن‌ها در یک نمودار رسم کنید.(شکل حاصل باید مجموع دو نمودار نرمال با میانگین و واریانس گفته شده در کنار هم باشد.)

۳. میانگین و واریانس هر دو توزیع را با استفاده از قانون اعداد بزرگ و بر حسب میانگین و واریانس نمونه‌ها تخمین بزنید و مقادیر بدست آمده را با مقادیر اصلی مقایسه کنید.

۴. با افزایش تعداد نمونه‌ها به ۱۰۰۰ و ۱۰۰۰۰ نتایج بدست آمده را گزارش کنید. آیا دقت مقادیر بدست آمده بهبود یافته است؟ توضیح دهید.

۵. در این قسمت تعداد نمونه‌های تولید شده در هر کلاس را برابر ۱۰۰۰ در نظر بگیرید. حال ۱۰۰ عدد داده‌ی تصادفی از یک توزیع نرمال با میانگین ۱.۵ و واریانس ۴ را تولید کنید. سپس با استفاده از یک طبقه‌بند نزدیکترین نمونه به میانگین [۱] و با استفاده از نرم یک [۲] داده‌ها را طبقه‌بندی کرده و نتایج را گزارش کنید.

۶. قسمت قبل را با نرم اقلیدسی نیز بررسی کنید.

۷. با استفاده از یک طبقه‌بند نزدیک‌ترین همسایه، داده‌های تولید شده در قسمت قبل را دوباره طبقه‌بندی کرده و نتایج را گزارش کنید.

---

[۱] Most Typical Sample
[۲] Manhatan Distance

۸. نتایج دو قسمت قبل را باهم مقایسه کرده و مزایا و معایب دو طبقه‌بند را گزارش کنید. برای بهبود معایب چه روشی پیشنهاد می‌کنید.

---

۲) (کدهای این سوال باید ضمیمه شود.)در این سوال قصد داریم تا سفید کردن داده‌ها را با استفاده از ماتریس کوواریانس انجام دهیم.

۱. ابتدا دو آرایه‌ی داده از یک توزیع نرمال، اولی با میانگین صفر و واریانس ۲۵ و دومی با میانگین ۱ و واریانس ۴ را تولید کنید. هر آرایه شامل ۱۰۰۰ داده باشد. سپس در یک نمودار دو بعدی آن‌ها را نمایش دهید.

۲. داده‌ها را با استفاده از ماتریس کوواریانس و مقادیر ویژه سفید کرده و داده‌های سفید شده را نمایش دهید.(روند سفیدسازی را با روابط مربوطه به صورت کامل توضیح دهید.)

---

۳) در این سوال قصد داریم تا الگوریتم بهینه‌سازی گرادیان کاهشی [۳] را بررسی کنیم.

۱. ابتدا الگوریتم گرادیان کاهشی را توضیح داده و روابط آن را برای تابع هزینه‌ی زیر بدست آورید.(پارامترهای θ، ماتریس وزن w و بردار بایاس b است.)

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{q}(h_\theta(x^{(i)}) - y^{(i)})^2$$

$$h(x) = \tanh(w^T x + b)$$

۲. قانون بروزرسانی وزن‌ها و بایاس‌ها را بدست آورده و اثر نرخ یادگیری [۴] را در قانون بروزرسانی وزن‌ها و بایاس توضیح دهید.(برای نرخ‌های یادگیری کم و بالا بررسی کنید.)

---

۴) تخمین Maximum Liklihood را با فرض مستقل بودن نمونه‌های آموزشی، برای پارامتر θ بدست آورید.

۱.

$$f(x_k, \theta) = \theta \exp(-\theta x_k) \qquad x_k \geq 0, \theta > 0$$

۲.

$$f(x_k, \theta) = \frac{x_k}{\theta^2} \exp(-\frac{x_k^2}{2\theta^2}) \qquad x_k \geq 0, \theta > 0$$

[۳] Gradient Decent
[۴] Learning Rate

.۳

$$f(x_k, \theta) = \sqrt{\theta} x_k^{\sqrt{\theta}-1} \qquad 0 \leq x_k \leq 1, \theta > 0$$

.۴

$$f(x_k, \theta) = \theta^2 x_k \exp(-\theta x_k) \qquad x_k > 0, \theta > 0$$

---

۵) (کدهای این سوال باید ضمیمه شود.)در این سوال می‌خواهیم سفیدسازی را روی داده‌های واقعی با ابعاد بالا انجام دهیم. ابتدا ۴ فایل بارگذاری شده همراه داده‌ها را دانلود کنید. سپس فایل train_data.csv را در برنامه خود بارگذاری کرده و در یک آرایه ذخیره کنید. (برای بارگذاری در پایتون می‌توانید از دستور "np.loadtxt('train_data.csv', delimiter=',', unpack=True)" استفاده کنید.) این فایل را در یک آرایه ذخیره کنید. آرایه باید شامل ۱۰۰۰۰ داده با ۴۰۰ ویژگی باشد.

پس از بارگذاری داده‌ها آن‌ها را با استفاده از ماتریس کوواریانس و مقادیر ویژه سفید کنید. میانگین و واریانس داده‌های سفید شده را بدست آورید.

---

۶) (کدهای این سوال باید ضمیمه شود.) در این سوال قصد داریم تا داده‌های بارگذاری شده را با استفاده از طبقه‌بند نزدیک‌ترین همسایه و طبقه‌بند نزدیک‌ترین نمونه به میانگین طبقه‌بندی کنیم.

۱. ابتدا ۴ فایل داده شده را که داده‌های آموزش شامل ۱۰۰۰۰ داده و داده‌های تست دارای ۱۰۰۰ داده است را در برنامه‌ی خود بارگذاری کنید.

۲. حال داده‌های تست را با استفاده از دو طبقه‌بند گفته شده، طبقه‌بندی کرده و دقت هر دو را باهم مقایسه کنید. همچنین دو طبقه‌بند گفته شده را از نظر زمان اجرا باهم مقایسه کنید. (دقت کنید که برای بدست آوردن میانگین دسته‌ها، از داده‌های آموزش استفاده کنید.)

۷) تفاوت سه فاصله‌ی نرم یک، نرم اقلیدسی و فاصله‌ی ماهالونوبیس [۵] را توضیح دهید. سپس قضیه‌ی زیر را برای دو بردار x و y اثبات کنید.(منظور از d_∞ نرم بی‌نهایت است.)

$$d_\infty(x, y) \leq d_2(x, y) \leq d_1(x, y)$$

---

[۵] Mahalonobis

۸) (سوال امتیازی) (کدهای این سوال باید ضمیمه شود.) ابتدا یک ماتریس ۳۰۰۰۰تایی که هر داده شامل ۲ ویژگی است، را از یک توزیع یکنواخت بین ۲- و ۲ تولید کنید. سپس به صورت تصادفی داده‌ها را به چهار دسته تقسیم کنید. حال الگوریتمی را برای سرعت بخشیدن به الگوریتم نزدیک‌ترین همسایه ارایه دهید.(سرعت الگوریتم پیشنهادی خود را با الگوریتم نزدیک‌ترین همسایه در حالت عادی، برحسب زمان طبقه‌بندی یک نمونه، مقایسه کنید.)

به نکات زیر توجه کنید:

- پاسخ شما باید در قالب یک فایل zip با عنوان شماره دانشجویی و نام و نام خانوادگی شما باشد.
- فایل ارسالی شما باید شامل یک فایل گزارش با فرمت pdf و یک پوشه حاوی کدهای مربوطه باشد. دقت کنید که کد مربوط به هر سوال و هر بخش باید به همان نام ذخیره شود.
- در صورت داشتن هرگونه سوال با دستیار آموزشی مربوطه از طریق ایمیل در ارتباط باشید.