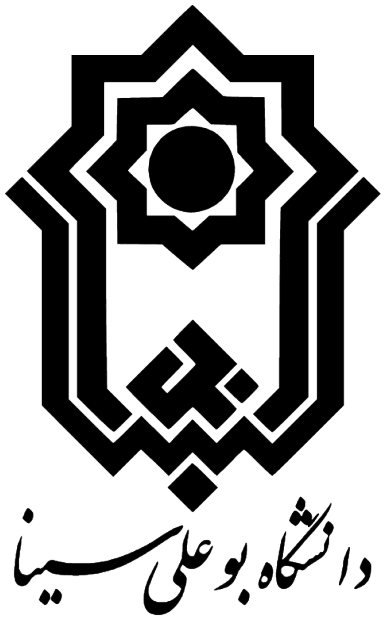


درس برنامه سازی پیشرفته

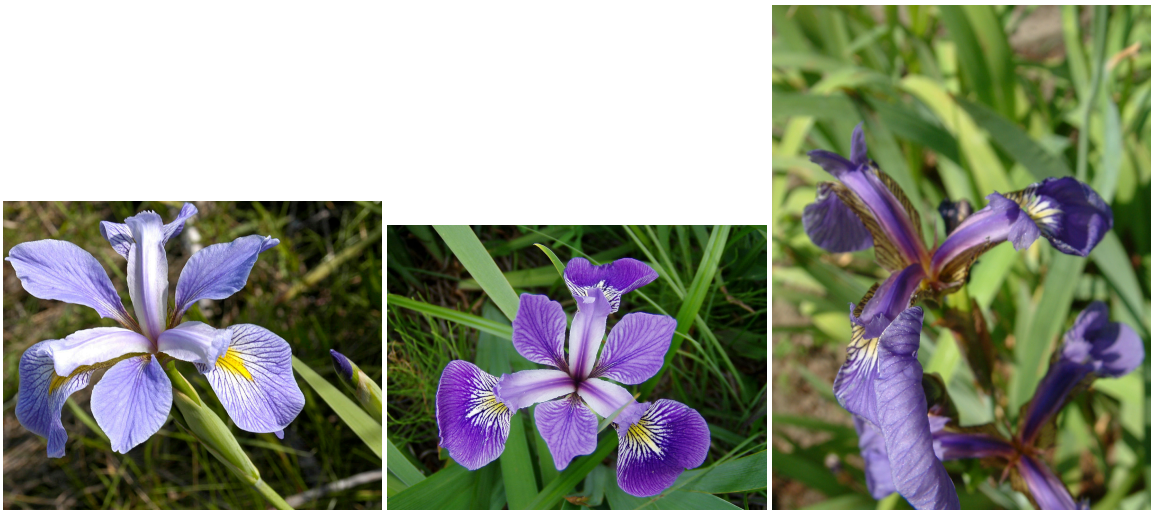
پیاده سازی MAP Classifier

زمستان ۹۷

شرح کلی

فرض کنید داده های پزشکی تعدادی از مراجعین به یک بیمارستان را در اختیار دارید. هر سطر این داده ها مربوط به یک بیمار است و شامل علائم پزشکی این فرد است. همچنین نتیجه‌ی آزمایش مبتلا بودن یا نبودن به یک سرطان خاص برای هر شخص وجود دارد است.حال تصور کنید که علائم پزشکی شخصی که در این لیست وجود ندارد را داشته باشیم. آیا می توان پیش بینی کرد که این شخص به سرطان مبتلا است یا نه؟ به عنوان یک سناریو دیگر در نظر بگیرید که داده های مربوط به سامانه‌ی تغذیه‌ی دانشگاه را در اختیار دارید. در این داده ها آمده است که هر شخص در چه زمانی چه غذایی را سفارش داده است. آیا بر اساس داده هایی که در اختیار داریم می توان پیشبینی کرد یک شخص در روز چهارشنبه‌ی این هفته از بین دو غذای سلف کدام یک را انتخاب خواهد کرد؟ آیا می توان قبول یا مردود شدن یک شخص در یک درس را براساس مشخصات و نمرات پیشین او و داده های دانشگاه پیش بینی کرد؟ در یادگیری ماشین به این دسته مسائل Classification یا رده‌بندی می گویند.در مثال بیمارستان می خواهیم از بین دو گزینه‌ی (ابتلا به سرطان) و (عدم ابتلا به سرطان) یکی را پیشبینی کنیم.در سناریوی سامانه‌ی تغذیه می خواهیم از بین گزینه های (جوجه کباب) و (خوراک مرغ) یکی را انتخاب کنیم.به این گزینه ها Class یا دسته یا رده می گویند.الگوریتم های زیادی برای Classification وجود دارد. در این تمرین قصد داریم یکی از روش های نسبتا ساده برای پاسخ به این مسائل را پیاده سازی کنیم.

شرح تمرین



از سمت راست Iris-Setosa و Iris-Versicolor و Iris-Virginica

هدف از این تمرین نوشتن یک برنامه است که یک فایل حاوی داده ها را دریافت و سپس یک مورد تست را با توجه به داده های داده شده به برنامه پیشبینی کند. با باز شدن برنامه ابتدا تعداد ویژگی ها یا feature های هر داده از کاربر پرسیده می شود.برای مثال داده های زیر مشخصات نمونه هایی از گل های Iris را نشان می دهد. این داده ها دارای ۴ ویژگی طول کاسبرگ، عرض کاسبرگ، طول گلبرگ و عرض گلبرگ برحسب سانتی متر هستند. بعد از اینکه کاربر تعداد ویژگی ها را وارد کرد بایستی تعداد رده ها یا گزینه ها یا کلاس ها از کاربر پرسیده شود. در داده

های زیر هر گل در یکی از سه رده Iris-setosa یا Iris-versicolor یا Iris-virginica دسته بندی می شود. پس تعداد رده ها در این مثال ۳ است.سپس نام این سه رده از کاربر پرسیده می شود. بعد از این باید کاربر نام فایل حاوی داده ها را وارد کند. فایل داده ها فایلی با شکل و فرمت زیر است. در سطر اول این فایل تعداد نمونه های موجود در فایل و در خطوط بعدی داده ها آمده است. در هر سطر ابتدا ویژگی ها و سپس رده‌ی مربوط به هر مورد نوشته شده است. این موارد با استفاده از Space از یکدیگر جدا شده اند.

```
7
5.1 3.5 1.4 0.2 Iris-setosa
4.9 3.0 1.4 0.2 Iris-setosa
4.7 3.2 1.3 0.2 Iris-setosa
4.6 3.1 1.5 0.2 Iris-setosa
5.7 2.8 4.1 1.3 Iris-versicolor
6.3 3.3 6.0 2.5 Iris-virginica
5.8 2.7 5.1 1.9 Iris-virginica
```

بعد از دریافت داده ها، کاربر تعداد داده های تست(داده هایی که رده‌ی آنها مشخص نیست) را وارد و برنامه آن ها را دریافت و رده‌ی آن ها را بر اساس داده های موجود و با استفاده از الگوریتمی که در قسمت بعد توضیح داده شده است، پیش بینی می کند. به همراه فایل توضیحات تمرین دیتاست اطلاعات گل ها پیوست شده است.می توانید برای تست برنامه‌ی خود از آن استفاده کنید.پنج داده‌ی زیر در دیتاست موجود نیستند.بعد از نوشتن برنامه می توانید عملکرد آن را با این پنج داده ارزیابی کنید:

```
5.2 4.1 1.5 0.1
5.5 4.2 1.4 0.2
6.0 2.9 4.5 1.5
5.7 2.6 3.5 1.0
6.3 2.8 5.1 1.5
```

رده‌ی این پنج داده به ترتیب به صورت زیر است:

```
Iris-setosa
Iris-setosa
Iris-versicolor
Iris-versicolor
Iris-virginica
```

می توانید بررسی کنید که برنامه می تواند رده‌ی این داده ها را پیش بینی کند یا نه.

شرح الگوریتم

در این تمرین می خواهیم پیشبینی رده‌ی داده ها بر اساس داده های موجود را با الگوریتمی که در ادامه توضیح خواهیم داد انجام دهیم. اساسا این الگوریتم بر اساس روشی به نام MAP یا Maximum a Posterior و با استفاده از قاعده‌ی بیز پیش بینی را انجام می دهد.فرض می کنیم تعداد K رده و تعداد n ویژگی و تعداد m نمونه داده داریم. اگر x_i نشان دهنده‌ی ویژگی شماره‌ی i باشد، احتمال اینکه یک داده به رده‌ی k تعلق داشته باشد متناسب است با :

$$p(C_k|x) = p(C_k) \times \prod_{i=1}^{n} p(x_i|C_k)$$

اگر N_k تعداد داده های متعلق به رده‌ی k باشد می توان $p(C_k)$ را به صورت زیر محاسبه کرد.

$$p(C_k) = \frac{N_k}{m}$$

مقدار $p(x_i|C_k)$ را باید برای هر ویژگی به صوت زیر حساب کرد:

$$p(x_i|C_k) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} e^{-\frac{(x_i-\mu_{ik})^2}{2\sigma_{ik}^2}}$$

به طوریکه x_i مقدار ویژگی i و μ_{ik} میانگین ویژگی i برای داده هایی که در رده‌ی k قرار دارند و σ_{ik}^2 واریانس ویژگی i برای داده هایی که در رده‌ی k قرار دارند است.

روال پیش بینی به این صورت است که با دریافت یک داده برای پیش بینی مقدار $p(C_k|x)$ را برای هر یک از رده ها محاسبه می کنیم.می توان گفت داده متعلق به رده ای است که این مقدار برای آن بیشینه است.

نکات مهم

- تعداد ویژگی ها توسط کاربر تعیین می شود و مقدار هر ویژگی عددی حقیقی است.
- از آن جایی که تعداد داده های برنامه می تواند زیاد باشد، روش مناسبی برای پاس دادن آرایه ها بین توابع استفاده کنید تا هزینه‌ی پاس دادن قابل قبول باشد.
- استفاده از امکانات Dynamic Memory Management زبان C++ مجاز است.
- استفاده از vector مجاز است.
- از متغیر های global استفاده نکنید.
- هیچ یک از توابع برنامه از جمله تابع main نباید بدنه ای طولانی داشته باشد.
- برای توسعه‌ی برنامه از Git استفاده کنید و کد برنامه‌ی خود را در یک Repository از نوع Private قرار دهید.
- برای برنامه Comment های مناسب و کافی بنویسید.