دانشکده فنی دانشگاه تهران

دانشکده برق و کامپیوتر

تمرین ۲ سیستم‌های هوشمند

طراحان:

میرحامد جعفرزاده

یاسمین نیکنام

رایانامه

Jafarzadeh.mirhamed@gmail.com

yasamin.1998@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_HW2_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. دیتاست زیر را در نظر بگیرید.

TOEFL	SOP	GPA	Research	Admission
Med	Yes	<8	No	No
High	Yes	>8	Yes	Yes
High	Yes	<8	No	Yes
Low	No	<8	No	No
Med	No	>8	Yes	No
High	Yes	>8	Yes	Yes
Med	No	>8	No	Yes
Low	No	>8	Yes	Yes
Low	Yes	<8	No	No

(آ) فرض کنید که Admission ویژگی است که میخواهیم آن را پیشبینی کنیم. در صورتی که از Information Gain استفاده کنیم، کدام یک از ویژگی ها را به عنوان ریشه ی درخت تصمیم با انشعاب های multi-way را بهتر است در نظر گرفت؟

(ب) فرض کنید که میخواهیم یک درخت تصمیم با انشعابهای دوتایی و با معیار Gini Index بسازیم. کدام یک از ترکیبهای نقطه انشعاب –ویژگی بهترین ترکیب برای Root Node است؟

- TOEFL-{ Low,Med }| { High }
- TOEFL-{ High, Low }| { Med }
- TOEFL-{ High, Med }| { Low }

۲. در این سوال، قصد داریم با پیاده‌سازی درخت تصمیم [۱] بر اساس الگوریتم ID3 ، داده‌های دیتاست [۲] prison_dataset.csv را طبقه‌بندی [۳] کنیم. ویژگی [۴] Recidivism - Return to Prison numeric را به عنوان هدف [۵] در نظر گرفته و می‌خواهیم با استفاده از ویژگی‌های دیگر، طبقه‌بندی را انجام دهیم.

(آ) با نمونه‌برداری تصادفی [۶] و به صورت ۸۰-۲۰ از دیتاست داده شده، داده‌ها را به داده‌های آموزش [۷] و آزمایش [۸] تقسیم کنید. با استفاده از الگوریتم ID3 درخت خود را پیاده‌سازی کنید و آن را با داده‌های آموزش، آموزش دهید. معیار انتخاب ویژگی برتر را Information Gain در نظر گرفته و عمق درخت خود را ۳ در نظر بگیرید. در نهایت لازم است دقت طبقه‌بند برای داده‌های آزمایش و همچنین Confusion Matrix را گزارش کنید.

(ب) حال قصد داریم برای بهبود عملکرد طبقه‌بند، از الگوریتم جنگل تصادفی [۹] استفاده کنیم. بدین منظور می‌توانید داده‌ها و ویژگی‌ها را تقسیم کرده و تعداد K درخت (حداقل ۳ درخت) را آموزش دهید و در نهایت با استفاده از Majority Voting ، دقت طبقه‌بند برای داده‌های آزمایش و همچنین Confusion Matrix را گزارش کنید. (دقت کنید که به شرط پیاده‌سازی درست، حتی بهبود در حد ۰.۱ درصد نسبت به قسمت قبل نیز قابل قبول است.)

(ج) در این قسمت با استفاده از کتابخانه‌ی Scikit-Learn ، الگوریتم جنگل تصادفی را برای random_state=0 max_depth=3 پیاده‌سازی کنید و دقت طبقه‌بند برای داده‌های آزمایش و همچنین Confusion Matrix را گزارش کرده و آن را با قسمت (ب) مقایسه کنید. دقت کنید به دلیل اینکه جنگل تصادفی مربوط به کتابخانه‌ی Scikit-Learn از مقادیر String برای ویژگی‌ها پشتیبانی نمی‌کند، لازم است که از LabelEncoder برای اینکار استفاده کنید. در این باره در اینترنت جستجو کنید و با استفاده از کتابخانه‌ی Scikit-Learn ، از روش مناسب برای رفع مشکل استفاده کنید.

---

[۱] Tree Decision
[۲] Dataset
[۳] Classify
[۴] Feature
[۵] Target
[۶] Sampling Random
[۷] Train
[۸] Test
[۹] Forest Random

۳. در این سوال، می‌خواهیم با استفاده از الگوریتم ژنتیک تعدادی کلمه‌ی رمزنگاری شده را بازیابی کنیم.
یکی از روش‌های مرسوم رمزنگاری، رمزنگاری به روش جایگزینی است. در این روش، هر حرف به یک حرف دیگر نگاشت داده می شود و در پیام اصلی جایگزین آن قرار داده می شود تا رمز موردنظر به دست آید. به این نگاشت ۲۶ حرفی کلید گفته می شود. جدول زیر نمونه اي از این کلید است.

Alphabet	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
Key	h	i	j	a	z	d	w	v	u	p	q	r	b	c	e	g	f	m	n	o	k	l	t	s	x	y

شکل ۱

برای حدس کلید صحیح که در مجموع ۲۶ حرف دارد، تمام حالات ممکن برای حدس کلید صحیح برابر ۲۶! است. واضح است که حدس چنین رمزی با تعداد حالات ممکن با روش های معمولی غیر ممکن است و می بایست از روش های فراابتکاری برای حل آن استفاده کرد.

نکات مهم جهت پیاده‌سازی:

- ابتدا باید جمعیت اولیه‌ای تعریف شود و هرکدام از اعضای آن بررسی شود. برای این کار به یک function fitting نیاز است که توسط شما انتخاب می‌شود.
- برای مشخص کردن میزان تناسب هرعضو، به یک لغت نامه نیاز است. برای این منظور، فایل دیکشنری برای شما آپلود شده است. این فایل شامل تعدادی کلمه که معنی دار محسوب می‌شوند، است.
- در ادامه، برای ایجاد جمعیت جدید نیاز به انجام crossover و mutation داریم. این عملیات تا زمانی انجام می‌شود که به کلید اصلی برسیم.

موفق باشید