بسم الله الرحمن الرحیم

دانشگاه علم و صنعت ایران

زمستان ۱۳۹۸

یادگیری ماشین　　تمرین سری چهارم　　تحویل: پنج‌شنبه ۲۹ اسفند

۱. فرض کنید در جامعه‌ای احتمال داشتن بیماری کرونا ۰,۱ و احتمال نداشتن بیماری ۰,۹ باشد. از ابزار تشخیص کرونا برای شناسایی افراد بیمار استفاده خواهد شد. خروجی این کیت اعداد صفر و یک می‌باشد. صفر به معنی نداشتن بیماری و یک به معنی مبتلا بودن است! اگر این کیت موارد مثبت را با دقت ۰,۹۶ و موارد منفی را با دقت ۰,۹۴ شناسایی کند و نتیجه آزمایش فردی مثبت باشد، احتمال مبتلا بودن و یا نبودن این فرد را به دست آورید.

۲. خطای بایاس به دلیل برخی فرضیات اشتباه در الگوریتم یادگیری پدید می‌آید. بایاس بالا، باعث عدم یادگیری ارتباط ورودی‌ها و خروجی‌ها می‌گردد (underfitting). واریانس نیز خطایی است که به دلیل حساسیت بالا نسبت به تغییرات کوچک در مجموعه داده‌ی آموزش به وجود می‌آید. این خطا باعث می‌شود تا الگوریتم داده‌های نویز را نیز به عنوان داده‌های اصلی مدل کند (overfitting). با توجه به توضیحات بالا مشخص نمایید کدام یک از الگوریتم‌های ذکر شده بایاس کم و کدام یک واریانس کمی دارند. سپس یک راه‌حل برای برطرف شدن ایراد الگوریتم ارائه دهید.

الف) رگرسیون خطی

ب) درخت تصمیم‌گیری

۳. فروش یک شرکت خودروسازی از سال ۱۳۹۴ تا ۱۳۹۸ در جدول زیر آمده است.

سال	۱۳۹۴	۱۳۹۵	۱۳۹۶	۱۳۹۷	۱۳۹۸
فروش	۱۲	۱۹	۲۹	۳۷	۴۵

الف) خط رگرسیونی که مربعات خطا را کمینه می‌کند به دست آورید. $(y = ax + b)$

ب) با استفاده از مدل به دست آمده در قسمت الف، فروش این شرکت در سال ۱۳۹۹ را تخمین بزنید.

ج) منحنی مرتبه دومی که مربعات خطا را کمینه می‌کند به دست آورید. $(y = ax^2 + bx + c)$

د) با استفاده از مدل به دست آمده در قسمت ج، فروش این شرکت در سال ۱۳۹۹ را تخمین بزنید.

۴. به کمک دانسته‌های خود از توزیع نرمال یک متغیره و با فرض مستقل بودن ویژگی‌ها، برای داده‌های موجود در فایل ضمیمه دسته‌بندهای ذکر شده را آموزش دهید و تحلیل نتایج را در گزارش بیاورید.

توجه: برای طراحی دسته‌بند از کدهای آماده استفاده نکنید. تنها می‌توانید برای توابع پایه ریاضی مانند میانگین و واریانس از کتابخانه‌های موجود استفاده کنید.