تمرین سری اول درس یادگیری ماشین

تاریخ ارسال: ۸ آبان ۹۶

تاریخ تحویل: ۲۳ آبان ۹۶

نمره: ۲ تا ۲,۵ نمره از ۲۰ نمره پایان ترم

فولدری به نام ML_Assignment01_name1_name2 بسازید که name1 و name2 نام افراد گروه هستند. در این فولدر سه فولدر به نام های ML_Problem11_name1_name2 و ML_Problem12_name1_name2 و ML_Problem13_name1_name2 بسازید و پاسخ سه مسئله را به ترتیب در این فولدرها قرار دهید. برای هر سوال پاسخ ها شامل کدها و فایل ورد (گزارش حل مسئله و جواب سوالات) است. دقت کنید در تمام فایل‌های کد و گزارش ها اسامی افراد تیم بصورت کامنت در بالای پاسخ‌ها عنوان شود. در نهایت فولدر اصلی را فشرده کنید و به آدرس taherian.khu@gmail.com ارسال کنید. فراموش نکنید که عبارت ML_Assignment01_name1_name2 (با جایگزینی نام افراد گروه) را در subject ایمیل بیاورید.

سوال اول: تغییرات آب و هوایی

مطالعات زیادی نشان داده که میانگین دما در قرن اخیر افزایش پیدا کرده است. در این سئوال رابطه بین میانگین دمای سراسری و بسیاری عوامل دیگر را بررسی می‌کنیم. فایل climate_change.csv داده‌های آب و هوایی از May 1983 تا December 2008 را در بر دارد. متغیرها به شرح زیر هستند:

- *Year*: the observation year.
- *Month*: the observation month.
- *Temp*: the difference in degrees Celsius between the average global temperature in that period and a reference value.
- *CO2*, *N2O*, *CH4*, *CFC.11*, *CFC.12*: atmospheric concentrations of carbon dioxide (CO_2), nitrous oxide (N_2O), methane (CH_4), trichlorofluoromethane (CCl_3F; commonly referred to as CFC-11) and dichlorodifluoromethane (CCl_2F_2; commonly referred to as CFC-12), respectively.
    - CO2, N2O and CH4 are expressed in ppmv (parts per million by volume -- i.e., 397 ppmv of CO2 means that CO2 constitutes 397 millionths of the total volume of the atmosphere)
    - CFC.11 and CFC.12 are expressed in ppbv (parts per billion by volume).
- *Aerosols*: the mean stratospheric aerosol optical depth at 550 nm. This variable is linked to volcanoes, as volcanic eruptions result in new particles being added to the atmosphere, which affect how much of the sun's energy is reflected back into space.
- *TSI*: the total solar irradiance (TSI) in W/m^2 (the rate at which the sun's energy is deposited per unit area). Due to sunspots and other solar phenomena, the amount of energy that is given off by the sun varies substantially with time.
- *MEI*: multivariate El Nino Southern Oscillation index (MEI), a measure of the strength of the El Nino/La Nina-Southern Oscillation (a weather effect in the Pacific Ocean that affects global temperatures).

الف- داده‌ها را به دو دسته traing (داده‌های تا آخر سال ۲۰۰۶) و test (داده‌های ابتدای ۲۰۰۷ به بعد) تقسیم کنید. از تمام خصیصه‌ها (متغیرهای مستقل) استفاده کنید و مدل رگرسیون بزنید تا Temp (متغیر وابسته) را پیش‌بینی کنید. توجه کنید که خصیصه‌های Year و Month نباید استفاده شوند. R^2 را محاسبه کنید. کدام خصیصه‌ها مهم تشخیص داده شده‌اند؟

ب- از داده‌های آموزشی استفاده کنید و همبستگی متغیرها را حساب کنید. با توجه به همبستگی بین خصیصه‌ها برخی را حذف کنید و مدل را با خصیصه‌های کمتر بسازید. R^2 را محاسبه کنید. کدام خصیصه‌ها مهم تشخیص داده شده‌اند؟

ج- در نظر گرفتن ترکیب‌های مختلفی از خصیصه‌ها و ساختن مدل برای هر ترکیب از خصیصه‌ها سخت و زمان‌بر است. تابع step این کار را برای شما انجام می‌دهد و ترکیبات مختلف از خصیه‌ها را بررسی می‌کند و در نهایت ترکیبی را در نظر می‌گیرد که هم R^2 خوبی داشته باشد و هم ساده باشد. یک مدل جدید با تابع step بسازید و نتایج را با مدل (الف) مقایسه کنید. (با استفاده از step? می‌توانید بیشتر راجع به این تابع بدانید.)

د- مدل به دست آمده از (ج) را روی داده‌های تست اعمال کنید و RMSE و R^2 را محاسبه کنید.

و- مدل را بهبود دهید. به عنوان مثال از درجه‌های بالاتر خصیصه‌ها استفاده کنید و یا regularization را اعمال کنید.

سوال دوم: پیش بینی بازپرداخت وام

LendingClub.com یک وبسایت است که ارتباط بین وام‌دهندگان و وام گیرندگان را برقرار می‌کند. در این سئوال ۹۵۷۸ داده مربوط به دریافت وام ۳ ساله توسط وام‌گیرندگان در این وبسایت را داریم (ماه می ۲۰۰۷ تا فوریه ۲۰۱۰). متغیر باینری not_fully_paid، متغیر وابسته (هدف) است که نشان می‌دهد وام گیرنده در بازپرداخت وام دچار مشکل شده یا خیر. متغیرهای مستقل که برای پیش‌بینی باید از آن‌ها استفاده کنید به شرح زیر هستند:

- **credit.policy**: 1 if the customer meets the credit underwriting criteria of LendingClub.com, and 0 otherwise.
- **purpose**: The purpose of the loan (takes values "credit_card", "debt_consolidation", "educational", "major_purchase", "small_business", and "all_other").
- **int.rate**: The interest rate of the loan, as a proportion (a rate of 11% would be stored as 0.11). Borrowers judged by LendingClub.com to be more risky are assigned higher interest rates.
- **installment**: The monthly installments ($) owed by the borrower if the loan is funded.
- **log.annual.inc**: The natural log of the self-reported annual income of the borrower.
- **dti**: The debt-to-income ratio of the borrower (amount of debt divided by annual income).
- **fico**: The FICO credit score of the borrower.
- **days.with.cr.line**: The number of days the borrower has had a credit line.
- **revol.bal**: The borrower's revolving balance (amount unpaid at the end of the credit card billing cycle).
- **revol.util**: The borrower's revolving line utilization rate (the amount of the credit line used relative to total credit available).
- **inq.last.6mths**: The borrower's number of inquiries by creditors in the last 6 months.
- **delinq.2yrs**: The number of times the borrower had been 30+ days past due on a payment in the past 2 years.
- **pub.rec**: The borrower's number of derogatory public records (bankruptcy filings, tax liens, or judgments).

الف- داده‌های دیتاست loans.csv را در یک dataframe بریزید و آن را کاوش کنید. چند درصد وام‌ها بازپرداخت شده‌اند؟ کدام خصیصه‌ها مقدار گم شده (missing value) دارند؟ چه روشی برای برخورد با داده‌های گم شده پیشنهاد می‌کنید؟

ب- ۷۰٪ داده‌ها را به عنوان داده training و بقیه را به عنوان داده test در نظر بگیرید. برای این کار ۷۰ درصد داده‌ها را به تصادف انتخاب کنید. سپس مدل logistic regressionروی داده‌های training بسازید. کدام خصیصه‌ها در این مدل مهم تشخیص داده شده‌اند؟

ج- نمودارهای پراکنش را رسم کنید. همبستگی خطی خصیصه‌ها را بررسی کنید. مدل‌های با خصیصه‌های کمتر بسازید و دقت را مقایسه کنید.

د- مدل به دست آمده از (ج) را روی داده‌های تست اعمال کنید. Confusion Matrix چیست؟ نمودار ROC را رسم کنید. مساحت زیر نمودار AUC(Area Under Curve) چقدر است؟ چه معیار دیگری برای ارزیابی کارایی مدل پیشنهاد می‌کنید؟

و- مدل را بهبود دهید. به عنوان مثال از درجه‌های بالاتر خصیصه‌ها استفاده کنید و یا regularization را اعمال کنید.

سوال سوم: نوشتن کد الگوریتم رگرسیون خطی

الف- تابعی linearRegressinFit را بنویسید که به عنوان ورودی X_train, y_train, alpha, threshold, max_iter, plotJ را بگیرد که alpha نرخ یادگیری است، max_iter حداکثر تعداد دفعات تکرار الگوریتم و threshold نشان دهنده معیار همگرایی الگوریتم است. تابع باید با استفاده از روش کاهش گرادیان ضرایب خط رگرسیون را پیدا کند. اگر در دو تکرار متوالی تفاضل تابع هزینه از threshold کمتر شود، الگوریتم باید متوقف شود (همگرا تشخیص داده شود). اگر تا max_iter تکرار به شرط همگرایی نرسد، الگوریتم متوقف شود و غیرهمگرا تشخیص داده شود. اگر plotJ، True باشد، باید در انتها نمودار J نسبت با شماره تکرار هم رسم شود. خروجی الگوریتم باید بردار ضرایب باشد. سعی کنید به جای استفاده از حلقه، از عملیات برداری و ماتریسی استفاده کنید تا سرعت بالا رود.

ب- تابعی به نام linearRegressionPredict بنویسید که X_test , coefficients را به عنوان ورودی بگیرد که coefficients بردار ضرایب مدل رگرسیون است. خروجی یک بردار (y_predict) است که مقادیر پیش بینی شده مدل برای داده‌های تست است.

ج- تابعی به نام evaluate بنویسید که y_test, y_predict را به عنوان ورودی بگیرد و بردار با طول ۲ برگرداند که اولین مولفه آن RMSE و دومین مولفه آن R^2 است.

د- تغییر کوچکی در تابع قسمت الف دهید به این ترتیب که یک ورودی باینری دیگر به نام regularization و landa هم بگیرد و اگر true بود، الگوریتم با regularization اعمال شود. Landa پارامتر مربوط به regularization است.