به نام خدا

یادگیری ماشین

تکلیف دوم– پاییز ۱۳۹۶ مهلت ارسال: ۱۷ آذر

قبل از اینکه شروع به حل تمرین‌ها کنید، حتما یک بار فایل Policies.pdf (موجود در وبگاه) که حاوی نکات مهم در تحویل تکالیف هست را مطالعه فرمایید.

حل سوالاتی که امتیازی هستند، اجباری نبوده و نمره‌ی اضافی خواهد داشت.

طراح هریک از سوالات در زیر مشخص شده است .در صورتی که سوالی دارید می توانید سوالتان را در گروه مطرح کنید یا از دستیار درس بپرسید.

طراحی سوالات:

۱-گلدانی ۲-گلدانی ۳- براساس تمرین CMU ۴-براساس تمرین UCI

---

۱. در این سوال به تخمین یک تابع سینوسی با استفاده از Polynomial Regression پرداخته می‌شود، سپس اثر regularization را برآن بررسی خواهد شد.

الف) تابع sin با ورودی زاویه که برحسب درجه بین ۳۰ تا ۳۰۰ درجه با فاصله ۵ درجه‌ای را درنظر بگیرید(۵، ۱۰، ۱۵،...) و مقداری نویز به آن بیافزایید.(تابعی تصادفی) سپس شکل داده ها را رسم نمایید.

ب) با استفاده از Polynomial regression تابع تخمینی با درجات ۱، ۳، ۵، ۷ و ۱۰ را به دست آورده و رسم نمایید. سپس پاسخ دهید ضرایب درجات مختلف چند جمله ای با زیاد شدن درجه چندجمله‌ای چه تغییری می‌کند.

ج) با استفاده از Regularization مقدار وزن را برای مقادیر مختلف λ به دست آورده و شکل حاصل از تخمین را برای چند جمله ای درجه ۷ رسم نمایید.

$$\text{Regularization}: \text{SSE} + \lambda \sum_i w_i^2$$

$$w = (X^T X + \lambda I)^{-1} X^T y$$

$$\lambda \in \{10^{-15}, 10^{-10}, 10^{-5}, 10^{-3}, 10^{-2}, 1, 5\}$$

د) با توجه به شکل‌هایی که در قسمت قبل رسم نموده اید تاثیر λ بر پیچیدگی مدل را شرح دهید.

۲. در صنعت برای ساخت یک بتن فاکتورهای مختلفی در نظر گرفته می‌شود و نیازمند ترکیبات مختلفی هستیم داده‌های این مساله براساس مقاله‌ای است که به بررسی این امر می‌پردازد.[1] قصد داریم با استفاده از linear regression و با استفاده از تکنیک Gradient descent بهترین خط برای تخمین را به دست آوریم. در اینجا تابع مجموع مربعات خطا را به عنوان cost function درنظر می‌گیریم. برای این منظور باید دو تابع Cost و Gradient را براساس مفاهیمی که فراگرفته اید بنویسید در این صورت تابع cost شما مقادیر x، y و W را به عنوان ورودی می‌گیرد و تابع Gradient شما باید مقادیر x، y، W، learning rate و Repetition (تعداد تکرار) را به عنوان ورودی بگیرد و مقادیر W و cost محاسبه شده در هر مرحله را بازگرداند.

$$h(x) = w_0 + w_1 x$$

$$J(w) = \frac{1}{2n}\sum_{i=1}^{n}(h_w(x_i) - y_i)^2$$

$$w_i \xleftarrow{\text{repeated until convergence}} w_i - \alpha \frac{\partial}{\partial w_i} J(w)$$

الف) با استفاده از روابط بالا تابع cost و Gradian را بنویسید.

ب) با استفاده از دو تابع مرحله قبل تابع فرضیه تخمین زننده روانی[2] براساس ویژگی مقدار سیمان[3] را برروی داده‌ها به دست آورید. و با شکل نشان دهید.(یک ویژگی و یک تابع هدف)

ج) تخمین بزنید اگر مقدار سیمان ۳۰۰ کیلوگرم باشد روانی چقدر خواهد بود.(آیا این تخمین منطقی است؟)

د) برای اینکه به تخمین بهتری دست یابیم، ویژگیهای دیگری را به ویژگی ها اضافه می‌کنیم همین مساله را به صورت multi variate linear regression برای تابع هدف روانی بتن حل کنید. و Wها را بدست آورید.

[1] Yeh, I-Cheng, "Simulation of concrete slump using neural networks," Construction Materials, Vol.162, No.1, 11-18, 2009
[2] FLOW
[3] Cement

۳. در این سوال با Overfitting و Regularized Logistic Regression آشنا خواهیم شد.

الف) تابع سیگموئید $\frac{1}{1+e^{-wX}}, X \in R$ برای وزن های افزایشی $w \in \{1,5,100\}$ رسم نمایید. از این نمودار برای تحلیل این مساله استفاده نمایید که چرا یک راه حل با اوزان بزرگ برای یک Logistic regression باعث ایجاد overfitting می‌شود.

ب) برای اینکه از overfitting جلوگیری نماییم. ما می‌خواهیم اوزان را کوچک نماییم. از این رو بجای محاسبه MLE به صورت زیر

$$\max_{w_0,...,w_d} \prod_{i=1}^{n} P(Y_i \mid X_i, w_0,..., w_d)$$

ما می‌خواهیم از MAP استفاده نماییم:

$$\max_{w_0,...,w_d} \prod_{i=1}^{n} P(Y_i \mid X_i, w_0,..., w_d) P(w_0,..., w_d)$$

در صورتی که $P(w_0,..., w_d)$ مقادیر Prior وزن ها می‌باشد.

برای بردار وزن‌ها یک توزیع گوسی استاندارد، N(0,I) درنظر بگیرید. برای به دست آوردن اوزان، اوزان بر اساس قاعده گرادیان صعودی[۴] باید هربار بروزرسانی شود این مرحله را نشان دهید.(راهنمایی: کافیست برای رابطه گرادیان صعودی عملگر تفریق در گرادیان نزولی را با جمع جایگزین نمایید)

[4] Ascent gradian

۴. (امتیازی) در این مساله قصد داریم linear regression را با استفاده از python شبیه سازی نماییم.

الف) داده های data را Load نمایید و این داده ها را به دو بخش train و test قسمت نمایید به اینصورت که یک چهارم آنها را به عنوان داده test و مابقی را به عنوان داده train درنظر بگیرید.

ب) برنامه ای بنویسید که خط حاصل از linear regression را رسم نماید.

ج) برنامه ای بنویسید که برای قسمت قبل MSE[5] را برای داده های train و test محاسبه نماید.

د) حال فرض نمایید که به مقدار هدف در داده های test قسمت قبل دسترسی ندارید و می‌خواهید بهترین درجه چند جمله ای برای تخمین را بیابید. چه راهی را پیشنهاد می‌کنید؟ برنامه آن را بنویسید.

[5] Mean Squared Error