بسمه تعالی

مقدمه ای بر یادگیری ماشین

نیمسال اول ۹۹-۹۸

تمرین عملی سری دوم

مدرس: صابر صالح

● مهلت تحویل تمرین ها: ۱۳۹۸/۰۸/۲۱ ●

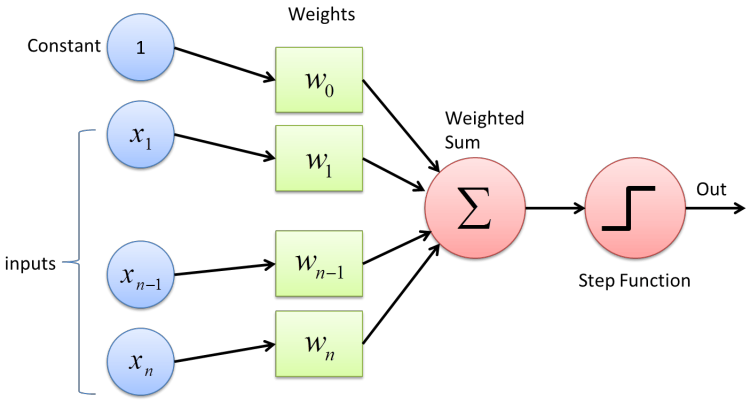
تمرین‌های برنامه‌نویسی

○ مقدمه

هدف از این تمرین پیاده سازی برخی الگوریتم های مهم یادگیری ماشین و آشنایی با روش های بهینه سازی عددی مانند گرادیان کاهشی می باشد. در بخش های اول و دوم الگوریتم های Perceptron و Logistic regression مورد بررسی قرار میگیرد و در بخش آخر الگوریتم Linear regression را پیاده سازی میکنید.

○ پرسپترون

به شبکه عصبی یک لایه perceptron می‌گویند. perceptron یک طبقه بند خطی باینری می‌باشد. شکل زیر نحوه عملکرد perceptron را نشان می‌دهد.



پیش بینی خروجی در پرسپترون از طریق رابطه زیر انجام می‌شود.

$$y' = \sum_{i=1}^{N} w_i x_i = w_1 x_1 + w_2 x_2 + ... + w_N x_N$$

خطای خروجی نیز به فرم زیر حساب می‌شود.

$$E = (y' - y)^2$$

که y مقدار واقعی خروجی می‌باشد.
الگوریتمی که در این تمرین برای به روز رسانی کردن وزن‌ها استفاده خواهد شد، الگوریتم گرادیان کاهشی می‌باشد. الگوریتم گرادیان کاهشی به فرم زیر می‌باشد.

$$\omega_{t+1} = \omega_t - \eta . \frac{dE}{d\omega} |_{\omega=\omega_t}$$

که η نرخ یادگیری است. برای تابع خطای معرفی شده در بالا، شکل ساده شده آن به صورت زیر می‌باشد.

$$\omega_{t+1} = \omega_t + \eta(y' - y)x$$

در روش Batch Mode برای آپدیت وزن‌ها از تمامی داده‌ها استفاده می‌شود اما در روش Online Mode با بررسی هر نمونه وزن‌ها آپدیت می‌شوند. در گزارش تمرین، این دو روش را از نظر بار محاسباتی و نرخ همگرایی مورد بررسی قرار دهید.

◦ Logistic regression

در این طبقه بند، خروجی با این عبارت مشخص میشود $P(Y=1|X=x_i,\omega)=\sigma(\omega\cdot x_i)$ که $\sigma(x)=\frac{1}{1+\exp(-x)}$ و y_i, x_i به ترتیب نمونه ی i ام ورودی و برچسب صفر و یا یک متناظر به آن میباشند. این طبقه بند در واقع توزیع شرطی برچسب خروجی بر حسب بردار ورودی را پیشبینی میکند. برای تابع هدف این الگوریتم از توابع MSE استفاده نشده و به جای آن از تابع Cross-Entropy استفاده میکنیم. در نهایت تابع هدف و گرادیان آن به صورت زیر بدست می آید. p_i همان احتمال بدست آمده برای داده x_i میباشد. توجه کنید که در رابطه Cross-Entropy فرض میشود توزیع q ، توزیع واقعی داده ها و توزیع p خروجی تابع sigmoid میباشد($q(Y=1|X=x_i)=y_i$). در گزارش تمرین ذکر کنید که چرا از این تابع به جایMSE برای تابع loss استفاده میشود.

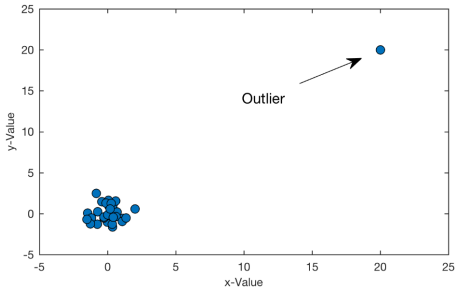
$$Cross-Entropy: H(q,p)=E_q[-\log p]$$

$$J(\omega)=-\frac{1}{N}\sum_{i=1}^{N}[y_i ln(p_i)+(1-y_i)ln(1-p_i)]$$

$$\nabla J(\omega)=-\frac{1}{N}\sum_{i=1}^{N}((p_i-y_i)x_i)$$

◦ داده های پرت

داده های پرت داده هایی هستند که فاصله ی قابل توجهی از توزیع داده های ورودی دارند. علت وجود این داده ها در مسائل یادگیری ماشین ناشی از خطاهایی است که به هنگام جمع آوردی داده ها به وجود آمده است و گاهی میتواند ناشی از توزیع خود داده ها نیز باشد. در هر دو صورت، حذف کردن این نوع داده از دیتاست میتواند بر روی دقت یادگیری و همگرایی سریع تر تاثیر مثبتی بگذارد. در شکل زیر مثالی از آن را مشاهده میکنید.



◦ Linear regression

این الگوریتم، برای تخمین توابعی که خروجی پیوسته دارند، مورد استفاده قرار میگیرد. در واقع هدف تخمین تابع $f: R^n \to R$ به صورت $f(x;\omega)=x\cdot\omega$ است. توابع هدف آن میتواند MSE و MAE و یا ترکیبی از هر دو باشد. در گزارش تمرین مزایا و معایب این دو تابع loss را نسبت به یکدیگر مقایسه کنید.

$$MSE: J(\omega)=\frac{1}{N}\sum_{i=1}^{N}(y_i-x_i\cdot\omega)^2$$

$$MAE: J(\omega)=\frac{1}{N}\sum_{i=1}^{N}|y_i-x_i\cdot\omega|$$

◦ تمارین

۱. طبقه بند Perceptron

دیتاست بخش پرسپترون شامل دو فایل classification_training.xlsx و classification_validation.xlsx می‌باشد که هر نمونه دارای ۹ ویژگی و برچسب گذاری باینری است. هدف این بخش پیاده سازی دو روش ذکر شده در بالا برای Gradient Descent برای آپدیت وزن‌های perceptron با استفاده از نرخ یادگیری های مختلف می‌باشد. $\eta=\{10^{-7},10^{-1}\}$

۲. طبقه بند Logistic regression

در این بخش دو ویژگی اول داده های Iris را جدا کرده و الگوریتم Logistic regression را بر روی آن اجرا کنید. دقت الگوریتم و نمودار داده ها به همراه خط بدست آمده توسط الگوریتم را رسم کنید. دقت کنید که تعدادی داده ی پرت در ورودی وجود دارد. یکبار الگوریتم را با حذف داده های پرت و بار دیگر با وجود آن ها رسم کنید. نمودار ها باید به ازای گام های زمانی مختلف و نرخ های یادگیری متفاوت رسم بشود.

۳. تمرین Linear regression

در این بخش الگوریتم Linear regression را بر روی داده های Buston housing اجرا کنید. داده ها شامل ۱۳ ویژگی و ۵۰۶ نمونه است که برای تخمین قیمت خانه ها به کار برده میشود. داده ها را به دو قسمت train و test با نسبت ۸۰ به ۲۰ تقسیم کنید. نمودار تابع Loss MSE را بر حسب نرخ های یادگیری متفاوت و تکرار های الگوریتم برای داده های train و test رسم کنید.

◦ در فایل starter_code.ipynb مربوطه، خواسته‌های مساله به صورت گام به گام ذکر شده است. نهایتا در گزارش تمرین، تحلیل خود را از نتایج حاصله بنویسید و سوال های بخش های قبل را پاسخ دهید. توجه کنید که برای پیاده سازی این الگوریتم ها مجاز به استفاده از کتابخانه های آماده مانند scikit learn نیستید.