یادگیری ماشین

نیم‌سال دوم ۱۴۰۲-۱۴۰۱

مدرس: دکتر سید ابوالفضل مطهری

دانشکده‌ی مهندسی کامپیوتر

تمرین سوم

۱. در عمل، $P_{\mathcal{X}\times\mathcal{Y}}$ عموما ناشناخته است و برای مدل‌کردن آن از کمینه‌کننده خطای تجربی [۱] (ERM) استفاده می‌کنیم. ما مسئله را به صورت یک مسئله رگرسیون خطی d بعدی بازنویسی می‌کنیم. در ابتدا در نظر داشته باشید که توابع درون $\mathcal{H}_d$ [۲] توسط بردار $\mathbf{b} = [b_0, b_1, \ldots, b_d]^T$ نمایش داده می‌شوند، به همین دلیل برای نمایش آن‌ها از نماد $f_{\mathbf{b}}$ استفاده می‌کنیم. به طرز مشابهی بردار $a \in \mathbb{R}$ [۳] برداری است که تابع $g(x) = f_{\mathbf{a}}(x)$ را نمایش می‌دهد. در نهایت داده‌های آموزشی را فرم ماتریسی به صورت مقابل نمایش می‌دهیم:

$$X = \begin{bmatrix} 1 & x_1 & \ldots & x_1^d \\ 1 & x_2 & \ldots & x_2^d \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_N & \ldots & x_N^d \end{bmatrix}, \quad \mathbf{y} = [y_0, y_1, \ldots, y_N]^T$$

استفاده از این نمادها اجازه استفاده از قضایا و تکنیک‌های جبرخطی را به ما می‌دهند. ماتریس X را ماتریس طراحی [۳] می‌نامیم.

- نشان دهید که کمینه‌کننده خطای تجربی (ERM) $\hat{\mathbf{b}}$ توسط کمینه‌سازی رابطه مقابل بدست می‌آید.

$$\hat{\mathbf{b}} = \arg\max_b ||Xb - \mathbf{y}||_2^2$$

- اگر $N > d$ باشد و ماتریس X دارای رتبه[۴] کامل باشد، نشان دهید $\hat{\mathbf{b}} = (X^TX)^{-1}X^T\mathbf{y}$. (راهنمایی: باید گرادیان تابع loss بالا را نسبت به $\mathbf{b}$ بگیرید). چرا شروط $N > d$ و کامل بودن رتبه ماتریس X لازم هستند؟

۲. $\mathbf{n}$ داده آموزش با $\mathbf{m}$ ویژگی[۵] فرض کنید. فرض کنید برچسب داده ها بردار $[y^{(0}, ..., y^{(n)}]$ و $X = [x^{(0)}, ...x^{(n)}]$ داده های آموزش باشند. ($X \in R^{n\times m}$).

- نشان دهید اگر یک مدل خطی تنها بر روی یکی از ویژگی‌ها آموزش دهیم، آنگاه $w_j = \frac{x_j^T y}{x_j^T x_j}$
- فرض کنید ستون‌های ماتریس X بر هم عمود [۶] باشند. نشان دهید پیدا کردن وزن‌های بهینه در این حالت با پیدا کردن هر وزن به صورت مستقل تفاوتی ندارد.

[1] Empirical Risk Minimizer
[2] Hypothesis Space (d dimension linear regression functions)
[3] Design Matrix
[4] Rank
[5] Feature
[6] Orthogonal

۳. فرض کنید $\hat{g}_۱$ و $\hat{g}_۲$ به صورت زیر تعریف شده اند.

$$\hat{g}_۱ = \underset{g}{\text{argmin}} = (\sum (y_i - g(x_i))^۲ + \lambda \int [g^{(۳)}(x)]^۲ dx)$$

$$\hat{g}_۲ = \underset{g}{\text{argmin}} = (\sum (y_i - g(x_i))^۲ + \lambda \int [g^{(۴)}(x)]^۲ dx)$$

توجه کنید که $g^{(m)}$ مشتق m ام تابع g است.

- اگر $\lambda \to \infty$ کدام یک از توابع RSS کمتری بر روی داده های آموزش دارند؟
- اگر $\lambda \to \infty$ کدام یک از توابع RSS کمتری بر روی داده های تست دارند؟

Spline

به موارد زیر پاسخ دهید:

۱. مزیت Spline نسبت به رگرسیون چندجمله‌ای با درجه آزادی برابر چیست؟

۲. تفاوت دو مورد Natural Spline و Restricted Cubic Spline را توضیح دهید.

۳. مفهوم Curse of Dimensionality یا نحسی ابعاد بالا را توضیح دهید و توضیح دهید این چه اثری بر روی استفاده از Spline و GAMها در ابعاد بالا خواهد گذاشت؟

GAM

به موارد زیر پاسخ دهید:

۱. در استفاده از Splineها تعداد knotها چطور انتخاب می‌شوند؟ یا به طور مشابه در GAMها تعداد توابع پایه را چطور انتخاب می‌کنید؟ مزایا و معایب تعداد بیشتر یا کمتر در هریک چیست؟ توضیح دهید.

۲. مزایا و معایب استفاده از GAMها در مقایسه با بقیه تکنیک های رگرسیون غیرخطی چیست؟ توضیح دهید.