به نام خدا

درس یادگیری ماشین
نیم‌سال دوم ۰۳-۰۴
استاد: ربیعی- دهقانیان

دانشکده مهندسی کامپیوتر

---

**آزمونک اول - ۱۴۰۳/۱۲/۱۹** زمان آزمون: ۳۰ دقیقه

---

۱. (۴۰ نمره) **(الف) گرادیان خطای میانگین مکعبی نسبت به $\theta$**

تابع هزینه داده شده به‌صورت زیر است:

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(\theta^T x^{(i)} - y^{(i)}\right)^3.$$

برای محاسبه‌ی گرادیان نسبت به $\theta$، ابتدا مشتق جزئی را حساب می‌کنیم:

$$\frac{\partial J}{\partial \theta} = \frac{1}{N}\sum_{i=1}^{N} 3\left(\theta^T x^{(i)} - y^{(i)}\right)^2 x^{(i)}.$$

بنابراین، گرادیان برابر است با:

$$\nabla_\theta J(\theta) = \frac{3}{N}\sum_{i=1}^{N}\left(\theta^T x^{(i)} - y^{(i)}\right)^2 x^{(i)}.$$

**(ب) حساسیت تابع هزینه نسبت به نقاط پرت**

تابع هزینه میانگین مربعات $\frac{1}{N}\sum(\theta^T x^{(i)} - y^{(i)})^2$ به دلیل وابستگی درجه‌ی دوم به خطا، تا حدی نسبت به نقاط پرت حساس است.

در مقابل، **تابع میانگین مکعبی حساسیت بیشتری دارد**، زیرا در آن خطا به توان سه می‌رسد. این بدان معناست که تأثیر خطاهای بزرگ (نقاط پرت) بسیار شدیدتر از MSE است، زیرا مقادیر بزرگ خطا هنگام مکعب شدن بسیار سریع رشد می‌کنند.

**(ج) مشکل اصلی استفاده از این تابع هزینه**

مشکل اصلی این تابع هزینه این است که **دارای گرادیان نامتقارن** است: خطاهای مثبت و منفی تأثیر یکسانی ندارند. به‌طور خاص، خطاهای مثبت مقدار بسیار بزرگ‌تری نسبت به خطاهای منفی در گرادیان ایجاد می‌کنند که باعث **عدم تعادل در یادگیری مدل** و رفتار ناپایدار در فرآیند بهینه‌سازی می‌شود.

۲. (۳۰ نمره) **(الف) آیا خطای دسته‌بندی روی داده‌های آموزشی می‌تواند افزایش یابد؟**

در حالت کلی، افزایش درجه‌ی $d$ باعث افزایش انعطاف‌پذیری مدل در نمایش داده‌ها می‌شود. اگر مدل توانایی جداسازی داده‌های آموزشی را داشته باشد، معمولاً انتظار داریم که **خطای آموزش کاهش یابد یا صفر شود**، زیرا مدل می‌تواند داده‌ها را با دقت بیشتری تطبیق دهد.

اما در مواردی خاص، ممکن است به دلیل مشکلات عددی (مانند افزایش هم‌خطی بین ویژگی‌های $(\phi_d(x))$، مدل دچار بی‌ثباتی در یادگیری شود که منجر به افزایش خطای آموزش می‌شود. با این حال، چنین حالتی معمولاً در عمل کمتر رخ می‌دهد.

بنابراین، در بیشتر موارد، انتظار داریم که **خطای آموزش کاهش یابد یا ثابت بماند**، اما در موارد خاص ممکن است به دلیل مشکلات عددی افزایش یابد.

**(ب) چرا خطای دسته‌بندی روی داده‌های جدید ابتدا کاهش و سپس افزایش می‌یابد؟**

در ابتدا، با افزایش $d$، مدل ویژگی‌های بیشتری از داده‌ها را در نظر می‌گیرد و می‌تواند بهتر الگوهای موجود را یاد بگیرد، بنابراین **خطای داده‌های جدید کاهش می‌یابد**. اما اگر $d$ بیش از حد افزایش یابد، مدل بیش از حد به داده‌های آموزشی می‌چسبد (پدیده‌ی **بیش‌برازش** یا *Overfitting*).

در این حالت، مدل نه‌تنها الگوهای کلی داده‌ها، بلکه نویز و جزئیات خاص مجموعه‌ی آموزشی را نیز یاد می‌گیرد. این موضوع باعث می‌شود که مدل روی داده‌های آموزشی عملکرد بسیار خوبی داشته باشد اما **روی داده‌های جدید دچار خطای بالا شود**، زیرا توانایی تعمیم‌دهی آن کاهش می‌یابد.

این رفتار مطابق با مفهوم **منحنی بایاس–واریانس** است که نشان می‌دهد افزایش پیچیدگی مدل ابتدا منجر به کاهش بایاس (افزایش دقت) می‌شود، اما پس از مدتی واریانس افزایش یافته و عملکرد روی داده‌های جدید بدتر می‌شود.

۳. (۳۰ نمره) **الف)** قرار می‌دهیم $Y_i = I[X_i \geq 0]$. در این صورت خواهیم‌داشت: ($L$ همان تابع Likelihood است)

$$L = \prod_{i=1}^{n} p^{Y_i}(1-p)^{1-Y_i}$$

$$\ln L = \sum_{i=1}^{n} Y_i \ln p + (1-Y_i)\ln(1-p)$$

حال گرادیان $\ln L$ نسبت به $p$ را برابر صفر قرار می‌دهیم تا پارامتر بهینه $p*$ به دست بیاید:

$$\frac{d \ln L}{dp*} = 0$$

$$\sum_{i=1}^{n} \frac{Y_i}{p*} - \frac{1-Y_i}{p*} = 0$$

$$\frac{\sum_{i=1}^{n} I[X_i \geq 0]}{p*} = \frac{n - \sum_{i=1}^{n} I[X_i \geq 0]}{1-p*}$$

$$p* = \frac{\sum_{i=1}^{n} I[X_i \geq 0]}{n}$$

**ب و ج)** توضیع پسین به شکل زیر خواهد بود: ($c$ یک مقدار ثابت است)

$$P(p \mid Data) = c \prod_{i=1}^{n} p^{Y_i}(1-p)^{1-Y_i} k p^{k-1}$$

$$\hat{p}_{MAP} = \arg\max_{p}\ \ln P(p \mid Data) = \arg\max_{p} \sum_{i=1}^{n} Y_i \ln p + (1-Y_i)\ln(1-p) + (k-1)\ln p$$

با مشتق گیری خواهیم‌داشت:

$$\frac{d \ln P(p \mid Data)}{dp*} = 0$$

$$\left(\sum_{i=1}^{n} \frac{Y_i}{p*} - \frac{1 - Y_i}{p*}\right) + \frac{k-1}{p*} = 0$$

$$\frac{\sum_{i=1}^{n} I[X_i \geq 0] + k - 1}{p*} = \frac{n - \sum_{i=1}^{n} I[X_i \geq 0]}{1 - p*}$$

$$p* = \frac{\sum_{i=1}^{n} I[X_i \geq 0] + k - 1}{n + k - 1}$$