جزوه‌ی جلسه اول: یادگیری ماشین (CE-۸۰۴۷) - رگرسیون خطی

تهیه‌شده توسط آرشیا قرونی و مهان بیغی

۲۱ سپتامبر ۲۰۲۴

فهرست مطالب

۱ مقدمه‌ای بر یادگیری ماشین

یادگیری ماشین شاخه‌ای از هوش مصنوعی است که به کامپیوترها امکان می‌دهد بدون برنامه‌نویسی صریح، از داده‌ها الگوهای معنادار استخراج کنند و عملکردشان را بهبود دهند. طبق تعریف کلاسیک Tom M. Mitchell:

> یک برنامه کامپیوتری از تجربه E نسبت به وظیفه T و معیار عملکرد P یاد می‌گیرد، اگر عملکردش در وظیفه T، که با P اندازه‌گیری می‌شود، با افزایش تجربه E بهبود یابد.

۱.۱ توضیح تکمیلی

یادگیری ماشین به ما کمک می‌کند تا به‌جای نوشتن قوانین پیچیده برای حل مسائل، به ماشین‌ها داده بدهیم و آن‌ها خودشان از این داده‌ها یاد بگیرند. مثلاً، به‌جای برنامه‌نویسی برای تشخیص ایمیل‌های اسپم، مدل یادگیری ماشین با دیدن مثال‌های ایمیل‌های اسپم و غیراسپم، خودش الگوها را پیدا می‌کند. این تعریف میچل به ما یادآوری می‌کند که یادگیری ماشین سه جزء اصلی دارد: وظیفه ،(Task) تجربه ،(Experience) و معیار عملکرد .(Performance)

۲.۱ کاربردهای یادگیری ماشین

- پیش‌بینی رفتار مشتریان (مانند تحلیل خرید)
- کنترل کیفیت در کارخانه‌ها
- تحلیل تصاویر پزشکی

۱.۲.۱ توضیح تکمیلی

کاربردهای یادگیری ماشین در دنیای واقعی بسیار گسترده‌اند. مثلاً، در تحلیل رفتار مشتریان، شرکت‌ها از داده‌های خرید برای پیشنهاد محصولات مناسب استفاده می‌کنند. در کارخانه‌ها، یادگیری ماشین می‌تواند نقص‌های محصولات را با تحلیل تصاویر شناسایی کند. در پزشکی، مدل‌ها می‌توانند با بررسی اسکن‌های MRI به تشخیص بیماری‌ها کمک کنند.

۲ یادگیری تحت نظارت

یادگیری تحت نظارت (Supervised Learning) شامل داده‌هایی است که هر نمونه شامل ورودی (x) و خروجی یا برچسب (y) است. هدف، یادگیری تابعی است که ورودی‌ها را به خروجی‌های درست نگاشت کند.

۱.۲ توضیح تکمیلی

یادگیری تحت نظارت مثل یادگیری با معلم است. داده‌های برچسب‌دار مثل تکالیفی هستند که جواب درستشون مشخصه. مثلاً، اگر بخواهیم قیمت خانه را پیش‌بینی کنیم، داده‌های ورودی می‌توانند متراژ، تعداد اتاق‌ها، و موقعیت جغرافیایی باشند، و خروجی قیمت واقعی خانه است. مدل باید یاد بگیرد که این ویژگی‌ها را به قیمت درست مرتبط کند.

۲.۲ انواع مسائل یادگیری تحت نظارت

- **رگرسیون**: پیش‌بینی مقادیر پیوسته (مثل قیمت خانه)
- **طبقه‌بندی**: پیش‌بینی دسته‌های گسسته (مثل تشخیص ایمیل اسپم)

۱.۲.۲ توضیح تکمیلی

در رگرسیون، خروجی یک عدد پیوسته است (مثل دما یا قیمت). در طبقه‌بندی، خروجی یک دسته یا کلاس است (مثل «اسپم» یا «غیراسپم»). انتخاب نوع مسئله به داده‌ها و هدف ما بستگی دارد. مثلاً، اگر بخواهیم پیش‌بینی کنیم که آیا فرد وام را بازپرداخت می‌کند یا نه، این یک مسئله طبقه‌بندی است.

۳ رگرسیون خطی

رگرسیون خطی یک مدل ساده برای پیش‌بینی مقادیر پیوسته است. فرضیه مدل به‌صورت زیر تعریف می‌شود:

$$h_w(x) = w_0 + w_1 x_1 + \cdots + w_D x_D = w^T x$$

که در آن:

- w: بردار وزن‌ها (پارامترهای مدل)
- x: بردار ویژگی‌های ورودی
- w_0: بایاس (عرض از مبدا)

۱.۳ توضیح تکمیلی

رگرسیون خطی فرض می‌کند که رابطه بین ورودی و خروجی خطی است. مثلاً، اگر بخواهیم قیمت خانه را پیش‌بینی کنیم، ممکن است فرض کنیم قیمت با متراژ (به‌صورت خطی) افزایش می‌یابد. w_0 به مدل اجازه می‌دهد که حتی اگر همه ویژگی‌ها صفر باشند، مقداری غیرصفر پیش‌بینی کند (مثل هزینه پایه خانه).

۲.۳ هدف رگرسیون خطی

یافتن بردار w که فاصله بین پیش‌بینی $(h_w(x))$ و مقدار واقعی (y) را کمینه کند.

۱.۲.۳ توضیح تکمیلی

هدف این است که پیش‌بینی‌های مدل تا حد ممکن به مقادیر واقعی نزدیک باشند. مثلاً، اگر مدل پیش‌بینی کند قیمت خانه ۱۰۰ میلیون است، اما قیمت واقعی ۱۱۰ میلیون باشد، خطای مدل ۱۰ میلیون است. ما می‌خواهیم این خطا را برای همه داده‌ها کم کنیم.

۴ تابع هزینه

تابع هزینه برای سنجش دقت مدل استفاده می‌شود. رایج‌ترین تابع هزینه، مجموع مربعات خطاها (SSE) است:

$$J(w) = \sum_{i=1}^{n} (y^{(i)} - h_w(x^{(i)}))^2$$

هدف، کمینه کردن $J(w)$ است تا مدل بهترین تطابق را با داده‌ها داشته باشد.

۱۰۴ توضیح تکمیلی

تابع هزینه مثل یک متر برای اندازه‌گیری کیفیت مدل عمل می‌کند. MSE (میانگین مربعات خطا) به این دلیل محبوب است که خطاهای بزرگ را بیشتر جریمه می‌کند (چون خطا را به توان ۲ می‌رساند). این باعث می‌شود مدل روی داده‌هایی که پیش‌بینی‌اش خیلی دور از واقعیت است، تمرکز بیشتری داشته باشد.

۵ حل تحلیلی (روش معادلات نرمال)

برای کمینه کردن تابع هزینه، می‌توان از مشتق‌گیری استفاده کرد و w را به‌صورت تحلیلی محاسبه کرد:

$$w = (X^T X)^{-1} X^T y$$

مزایا:

- دقیق و بدون نیاز به تکرار

محدودیت‌ها:

- محاسبات سنگین برای داده‌های بزرگ
- نیاز به معکوس‌پذیری ماتریس $X^T X$

۱۰۵ توضیح تکمیلی

روش معادلات نرمال مثل حل یک معادله ریاضی است که جواب دقیق می‌دهد. اما وقتی تعداد داده‌ها زیاد باشد (مثلاً میلیون‌ها نمونه)، محاسبه معکوس ماتریس $X^T X$ خیلی زمان‌بر و پرهزینه است. همچنین، اگر داده‌ها هم‌خطی باشند (یعنی برخی ویژگی‌ها خیلی شبیه هم باشند)، ماتریس معکوس‌پذیر نیست.

۶ گرادیان نزولی

گرادیان نزولی (Gradient Descent) یک روش عددی برای کمینه کردن تابع هزینه است. در هر مرحله، وزن‌ها به‌روزرسانی می‌شوند:

$$w_{t+1} = w_t - \eta \nabla J(w_t)$$

که در آن:

- η: نرخ یادگیری (Learning Rate)
- $\nabla J(w_t)$: گرادیان تابع هزینه

۱۰۶ توضیح تکمیلی

گرادیان نزولی مثل این است که در یک کوه بخواهیم پایین‌ترین نقطه را پیدا کنیم. گرادیان به ما جهت شیب را نشان می‌دهد، و نرخ یادگیری تعیین می‌کند که قدم‌های ما چقدر بزرگ باشند. اگر قدم‌ها خیلی بزرگ باشند، ممکن است از نقطه بهینه رد شویم؛ اگر خیلی کوچک باشند، رسیدن به جواب طول می‌کشد.

۲۰۶ انواع گرادیان نزولی

- **GD Batch**: استفاده از کل داده‌ها در هر مرحله (دقیق اما کند)
- **GD Stochastic**: استفاده از یک نمونه در هر مرحله (سریع اما ناپایدار)
- **GD Mini-batch**: استفاده از زیرمجموعه‌ای از داده‌ها (تعادل بین سرعت و دقت)

۱۰۲۰۶ توضیح تکمیلی

GD Batch مثل این است که کل نقشه کوه را یکجا ببینیم و قدم برداریم، ولی محاسبه‌اش سنگین است. Stochastic GD مثل این است که فقط به یک نقطه نگاه کنیم و سریع حرکت کنیم، ولی ممکن است مسیر پرنوسانی طی کنیم. GD Mini-batch یک تعادل خوب است و در عمل (مثل یادگیری عمیق) خیلی استفاده می‌شود.

۳۰۶ نکات کلیدی گرادیان نزولی

- انتخاب η مناسب حیاتی است: خیلی بزرگ واگرایی، خیلی کوچک کندی
- نرمال‌سازی داده‌ها باعث بهبود همگرایی می‌شود

۱۰۳۰۶ توضیح تکمیلی

نرمال‌سازی داده‌ها (مثل استاندارد کردن ویژگی‌ها به میانگین صفر و واریانس یک) باعث می‌شود گرادیان‌ها در مقیاس مشابه باشند و الگوریتم سریع‌تر همگرا شود. همچنین، تکنیک‌هایی مثل کاهش تدریجی نرخ یادگیری یا استفاده از روش‌های پیشرفته‌تر (مثل Adam) می‌توانند عملکرد را بهبود دهند.

۷ رگرسیون چندجمله‌ای

وقتی رابطه بین ورودی و خروجی غیرخطی است، از رگرسیون چندجمله‌ای استفاده می‌شود. فرضیه مدل به‌صورت زیر است:

$$h(x) = w_0 + w_1 x + w_2 x^2 + \cdots + w_m x^m$$

این روش مشابه رگرسیون خطی است، اما ویژگی‌ها به‌صورت توان‌های مختلف تبدیل می‌شوند.

۱۰۷ توضیح تکمیلی

رگرسیون چندجمله‌ای به ما اجازه می‌دهد روابط پیچیده‌تر (مثل منحنی‌ها) را مدل کنیم. مثلاً، اگر بخواهیم فروش یک محصول را بر اساس دمای هوا پیش‌بینی کنیم، ممکن است فروش در دماهای میانی بیشتر باشد و در دماهای خیلی بالا یا پایین کمتر شود. این رابطه غیرخطی را می‌توان با یک چندجمله‌ای مدل کرد.

۲۰۷ مزایا و معایب

- **مزیت**: مدل‌سازی روابط پیچیده‌تر
- **معایب**: خطر بیش‌برازش (Overfitting) در صورت استفاده از درجه‌های بالا

۱۰۲۰۷ توضیح تکمیلی

بیش‌برازش وقتی رخ می‌دهد که مدل نه‌تنها الگوهای واقعی، بلکه نویزهای داده را هم یاد می‌گیرد. مثلاً، اگر یک چندجمله‌ای درجه ۱۰ برای ۱۰ داده فیت کنیم، مدل ممکن است دقیقاً از همه نقاط بگذرد، اما برای داده‌های جدید عملکرد ضعیفی داشته باشد. برای جلوگیری از این مشکل، از تکنیک‌هایی مثل تنظیم‌سازی (Regularization) یا انتخاب مدل با داده‌های اعتبارسنجی استفاده می‌شود.

۸ نکات کلیدی برای یادگیری و مصاحبه

- **Underfitting**: وقتی مدل بیش‌ازحد ساده است و داده‌ها را خوب مدل نمی‌کند.
- **Overfitting**: وقتی مدل بیش‌ازحد پیچیده است و نویز داده‌ها را هم یاد می‌گیرد.
- **راه‌حل**: استفاده از داده‌های اعتبارسنجی (Validation Data) و تکنیک‌هایی مثل تنظیم‌سازی (Regularization).

۱۰۸ توضیح تکمیلی

Underfitting مثل این است که بخواهیم یک منحنی پیچیده را با یک خط صاف مدل کنیم؛ مدل نمی‌تواند الگوهای داده را خوب یاد بگیرد. Overfitting برعکس، مثل این است که مدل بیش‌ازحد به داده‌های آموزشی وابسته شود و برای داده‌های جدید خوب کار نکند. داده‌های اعتبارسنجی به ما کمک می‌کنند تا مدلی را انتخاب کنیم که نه خیلی ساده باشد و نه خیلی پیچیده.

۲۰۸ جدول سوالات مصاحبه

مفهوم	سوال احتمالی در مصاحبه
رگرسیون خطی	تفاوت بین روش تحلیلی و گرادیان نزولی چیست؟
تابع هزینه	چرا از MSE به‌عنوان تابع هزینه استفاده می‌کنیم؟
بیش‌برازش	چگونه می‌توان از بیش‌برازش جلوگیری کرد؟
گرادیان نزولی	تفاوت Stochastic Batch، و Mini-batch چیست؟
معادلات نرمال	محدودیت‌های روش Pseudo-Inverse چیست؟

۱۰۲۰۸ توضیح تکمیلی

این سوالات معمولاً در مصاحبه‌های فنی یادگیری ماشین پرسیده می‌شوند. مثلاً، برای سوال «چرا MSE؟»، می‌توانید بگویید که MSE خطاهای بزرگ را بیشتر جریمه می‌کند و محاسباتش ساده است. برای جلوگیری از بیش‌برازش، تکنیک‌هایی مثل L۱/L۲ Regularization، Dropout (در شبکه‌های عصبی)، یا انتخاب مدل با Cross-Validation را ذکر کنید.

۹ جمع‌بندی

رگرسیون خطی یکی از پایه‌ای‌ترین مدل‌های یادگیری ماشین است که برای پیش‌بینی مقادیر پیوسته استفاده می‌شود. با درک مفاهیم تابع هزینه، روش‌های بهینه‌سازی (تحلیلی و گرادیان نزولی)، و رگرسیون چندجمله‌ای، می‌توانید پایه‌ای محکم برای یادگیری سایر الگوریتم‌های یادگیری ماشین ایجاد کنید.

۱۰۹ توضیح تکمیلی

این جلسه یک نقطه شروع عالی برای یادگیری ماشین است، چون رگرسیون خطی مفاهیم اصلی مثل تابع هزینه، بهینه‌سازی، و تعادل بین پیچیدگی و دقت را به‌خوبی معرفی می‌کند. این مفاهیم در الگوریتم‌های پیچیده‌تر مثل شبکه‌های عصبی یا SVM هم کاربرد دارند.