تمرین 2 شبکه های عصبی – مازندرانیان - 830402066

سوال 1) روش های انتخاب بهترین مقدار Learning Rate(α) را معرفی کنید.

روش‌های هیورستیک

روش‌های هیورستیک، روش‌هایی هستند که بر اساس تجربه و آزمون و خطا انجام می‌شوند. این روش‌ها معمولاً سریع‌تر هستند اما ممکن است به نتیجه مطلوب نرسند..

شروع با یک مقدار کوچک: معمولاً با یک مقدار کوچک مانند 0.01 یا 0.001 شروع می‌شود و سپس به صورت تدریجی افزایش داده می‌شود.

استفاده از یک بازه: یک بازه از مقادیر ممکن برای نرخ یادگیری را در نظر بگیرید و با آزمایش هر مقدار، بهترین نتیجه را انتخاب کنید.

کاهش نرخ یادگیری به صورت تدریجی: با پیشرفت آموزش، نرخ یادگیری را به تدریج کاهش دهید تا مدل به یک مینیمم محلی همگرا شود.

روش‌های مبتنی بر گرادیان

این روش‌ها از اطلاعات گرادیان برای تنظیم نرخ یادگیری استفاده می‌کنند و معمولاً نتایج بهتری نسبت به روش‌های هیورستیک ارائه می‌دهند.

Adam: یکی از محبوب‌ترین الگوریتم‌های بهینه‌سازی است که به صورت خودکار نرخ یادگیری را برای هر پارامتر تنظیم می‌کند. این الگوریتم از یک تخمین تطبیقی از لحظه اول و دوم گرادیان استفاده می‌کند.

RMSprop: این الگوریتم مشابه آدام است اما از یک میانگین موزون از مربع‌های گرادیان استفاده می‌کند.

Adagrad: این الگوریتم برای هر پارامتر یک نرخ یادگیری جداگانه در نظر می‌گیرد.

SGD با اندازه گام متغیر: در این روش، مقدار نرخ یادگیری در هر مرحله بر اساس شیب تابع هزینه به روز می‌شود.

AdaGrad

$$g_0 = 0$$

$$g_{t+1} \leftarrow g_t + \nabla_\theta \mathcal{L}(\theta)^2$$

$$\theta_j \leftarrow \theta_j - \epsilon \frac{\nabla_\theta \mathcal{L}}{\sqrt{g_{t+1}} + 1e^{-5}}$$

RMS Prop

$$g_0 = 0\,, \alpha \simeq 0.9$$

$$g_{t+1} \leftarrow \alpha \cdot g_t + (1-\alpha)\nabla_\theta \mathcal{L}(\theta)^2$$

$$\theta_j \leftarrow \theta_j - \epsilon \frac{\nabla_\theta \mathcal{L}}{\sqrt{g_{t+1}} + 1e^{-5}}$$

پارامترهای پیکربندی الگوریتم بهینه سازی آدام (منبع: فرادرس)

- **alpha:** با عنوان «نرخ یادگیری» یا «طول گام» نیز از آن یاد می‌شود. آلفا نسبتی است که وزن‌ها بر اساس آن به‌روزرسانی می‌شوند (به عنوان مثال ۰.۰۰۱). مقادیر بزرگ‌تر برای پارامتر alpha (مثلاً ۰.۳) باعث می‌شود یادگیری اولیه سریع‌تر و پیش از آن انجام شود که میزان نرخ مربوطه به‌روزرسانی می‌شود. مقادیر کوچک‌تر برای پارامتر آلفا (به عنوان مثال 1.0E-5)در طول آموزش، سرعت یادگیری را کاهش می‌دهد.
- **beta1:** نرخ فروپاشی نمایی برای تخمین‌های گشتاور اول است (مثلاً مقدار آن می‌تواند ۰.۹ باشد).
- **beta2 :**نرخ فروپاشی نمایی برای تخمین‌های لحظه دوم است (مثلاً مقدار آن می‌تواند ۰.۹ باشد). این مقدار باید در مسائلی با گرادیان تُنُک) مانند NLP و مسائل بینایی کامپیوتر (نزدیک به یک باشد.
- **epsilon :**عددی بسیار کوچکی مانند (8 - 10e) است که در پیاده‌سازی‌ها برای جلوگیری از تقسیم بر صفر استفاده می‌شود.

الگوریتم ADAM ترکیبی از الگوریتم های Momentum و RMSprop است. Momentum به پارامترها اجازه می‌دهد تا در جهت گرادیان، شتاب بگیرند. این کار باعث می‌شود که پارامترها بتوانند سریع‌تر به سمت مینیمم محلی حرکت کنند و از گیر افتادن در مینیمم‌های محلی کوچک جلوگیری کنند. RMSprop به صورت تطبیقی نرخ یادگیری را برای هر پارامتر تنظیم می‌کند. این کار باعث می‌شود که پارامترهایی که گرادیان‌های بزرگ‌تری دارند، با سرعت کمتری به‌روزرسانی شوند و پارامترهایی که گرادیان‌های کوچکتری دارند، با سرعت بیشتری به‌روزرسانی شوند. Adam از دو متغیر کمکی استفاده می‌کند: یکی برای محاسبه میانگین نمایی گرادیان‌ها (مشابه momentum) و دیگری برای محاسبه میانگین نمایی مربعات گرادیان‌ها (مشابه rmsprop) سپس از این دو متغیر برای به‌روزرسانی پارامترها استفاده می‌شود.

$$v_t = \gamma * v_{t-1} + \eta * g_t$$
$$\theta = \theta - v_t$$

الگوریتم momentum

$$E[g_t^2] = \gamma * E[g_{t-1}^2] + (1 - \gamma) * g_t^2$$
$$\theta = \theta - (\eta / \sqrt{(E[g_t^2] + \varepsilon)}) * g_t$$

الگوریتم RMSProps

$$m_t = \beta_1 * m_{t-1} + (1 - \beta_1) * g_t$$
$$v_t = \beta_2 * v_{t-1} + (1 - \beta_2) * g_t^2$$
$$\theta = \theta - (\alpha * m_t / \sqrt{(v_t + \varepsilon)})$$

الگوریتم ADAM

البته الگوریتم ADAM در بعضی نواحی به یک راه حل بهینه همگرا نمی شود به همین دلیل در برخی مسائل از الگوریتم هایی مثل گرادیان کاهش تصادفی به همراه گشتاور استفاده می شود.

گشتاور N ام یک متغیر تصادفی به عنوان مقدار مورد انتظار آن متغیر به توان n تعریف می‌شود. رابطه آن در ادامه آمده است:

$$mn = E[Xn]mn = E[Xn]$$

در فرمول بالا، m نماد گشتاور یا moment و X نماد متغیر تصادفی است. گرادیان تابع هزینه شبکه عصبی را می‌توان به عنوان یک متغیر تصادفی در نظر گرفت. زیرا معمولاً روی برخی از دسته‌های تصادفی کوچکی از داده‌ها ارزیابی می‌شود. مومنت اول میانگین است و مومنت دوم واریانس غیر مرکزی (به این معنی که میانگین در طول محاسبه واریانس کم نمی‌شود) محسوب می‌شود.

سوال 2) L2 Regularization چه نقشی در انتخاب ویژگی ها دارد؟

این الگوریتم با اضافه کردن یک جریمه وزن مدل ها را به صفر میل می دهد که باعث میشود مدل به ویژگی های کمتری وابسته شود و پیچیدگی آن کم شود.

مثال از ChatGPT:

فرض کنید می‌خواهیم یک مدل رگرسیون خطی برای پیش‌بینی قیمت خانه بسازیم. ویژگی‌های ورودی ما عبارتند از:

- مساحت خانه (x1)
- تعداد اتاق‌ها (x2)
- سن خانه (x3)
- فاصله تا مرکز شهر (x4)
- وجود پارکینگ (x5 اگر پارکینگ داشته باشد 1 در غیر این صورت 0)

$$y = w0 + w1 * x1 + w2 * x2 + w3 * x3 + w4 * x4 + w5 * x5$$

در اینجا:

قیمت خانه : Y

وزن‌های مدل :$w0, w1, w2, w3, w4, w5$

ویژگی‌های ورودی :$x1, x2, x3, x4, x5$

اکنون L2 Regularization را به تابع هزینه اضافه می‌کنیم:

$$J(w) = (1/2m) * \Sigma(y\_i - y\_hat\_i)\text{^}2 + \lambda * \Sigma(w\_i\text{^}2)$$

فرض کنید بعد از آموزش مدل با L2 Regularization، وزن‌های زیر حاصل شود:

$$w1 = 0.5$$

$$w2 = 0.3$$

$$w3 = 0.01$$

$$w4 = 0.2$$

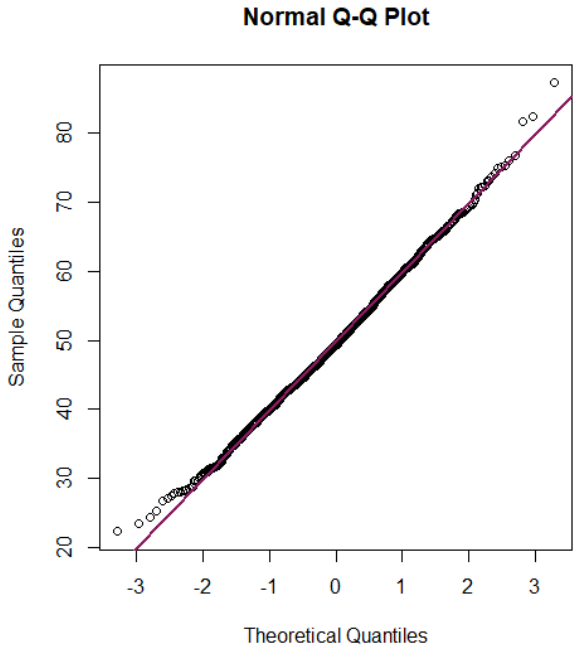
$$w5 = 0.4$$

در این حالت، وزن ویژگی سن خانه (w3) بسیار کوچک است. این نشان می‌دهد که سن خانه تأثیر کمی بر قیمت خانه دارد. بنابراین، می‌توان این ویژگی را از مدل حذف کرد و مدل ساده‌تری ساخت.

سوال 3) چطور یک آزمون فرض آماری طرح کنیم که بفهمیمم توزی آماری نرمال برای دیتای ما مناسب است؟

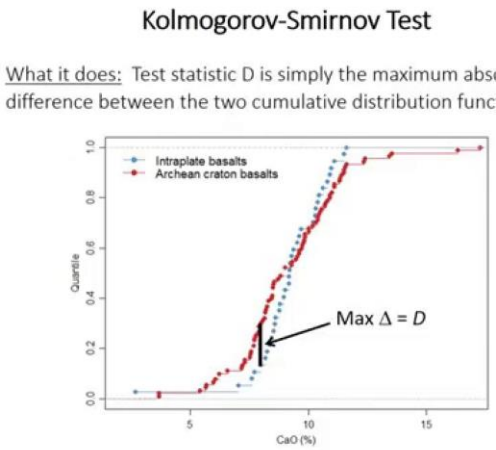
1. روش‌های گرافیکی:

- هیستوگرام :با رسم هیستوگرام می‌توانیم شکل کلی توزیع داده‌ها را مشاهده کنیم. اگر توزیع به شکل زنگوله‌ای باشد، احتمالاً داده‌ها نرمال هستند.
- نمودار Q-Q: این نمودار مقادیر مشاهده شده را در مقابل مقادیر مورد انتظار در یک توزیع نرمال رسم می‌کند. اگر نقاط روی یک خط مستقیم قرار بگیرند، داده‌ها به احتمال زیاد نرمال هستند.



2. آزمون‌های آماری:

- آزمون کولموگروف-اسمیرنوف :این آزمون فاصله بین توزیع تجمعی مشاهده شده و توزیع تجمعی نرمال را محاسبه می‌کند. اگر این فاصله کوچک باشد، داده‌ها به احتمال زیاد نرمال هستند.
- آزمون شاپیرو-ویلک :این آزمون به طور خاص برای نمونه‌های کوچک طراحی شده است و به حساسیت بیشتری نسبت به انحراف از نرمال بودن دارد.

[redacted]


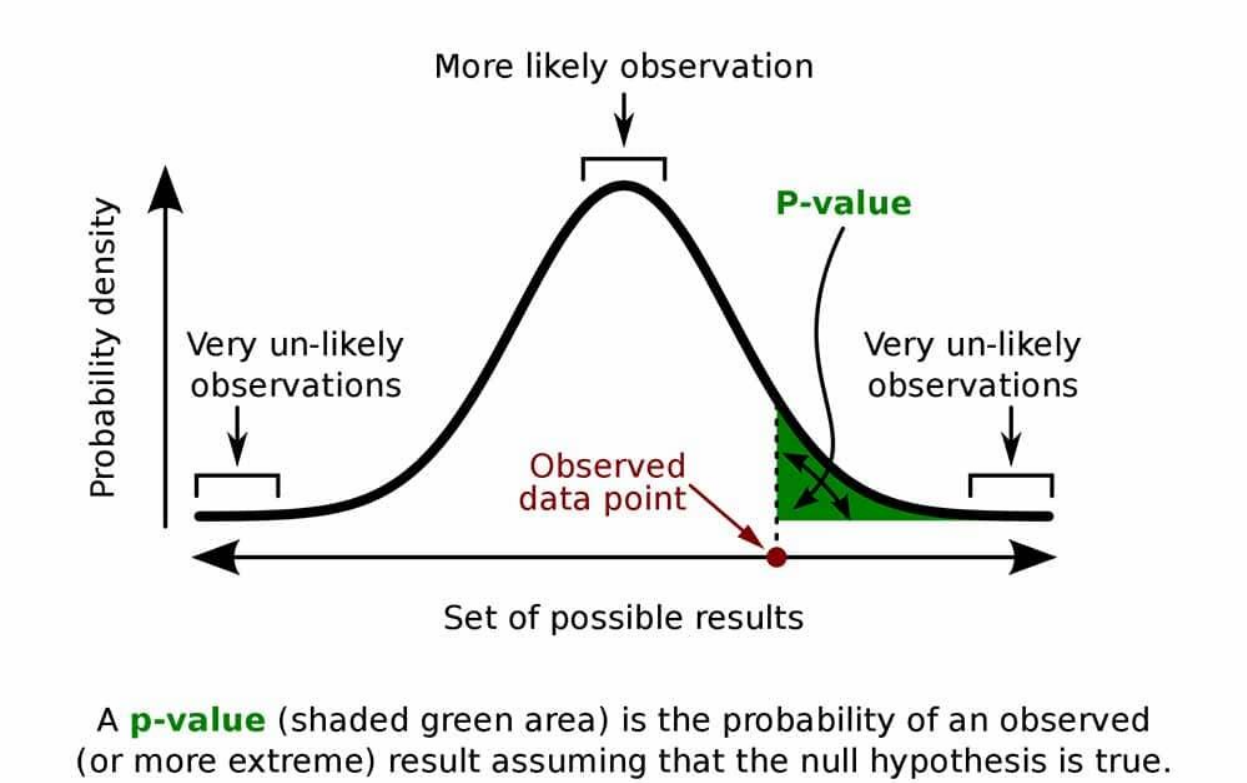
[redacted]

مراحل انجام آزمون فرض آماری برای نرمال بودن داده‌ها:

1. طرح فرضیه‌ها :

- فرض صفر :(H0) داده‌ها از یک توزیع نرمال پیروی می‌کنند.
- فرض مقابل :(H1) داده‌ها از یک توزیع نرمال پیروی نمی‌کنند.

2. انتخاب سطح معنی‌داری :(α) معمولاً سطح معنی‌داری 0.05 انتخاب می‌شود.
3. انتخاب آزمون مناسب :با توجه به حجم نمونه و نوع داده‌ها، آزمون کولموگروف-اسمیرنوف یا شاپیرو-ویلک را انتخاب کنید.
4. محاسبه آماره آزمون :نرم‌افزارهای آماری مانندSPSS ، SAS و R به طور خودکار آماره آزمون را محاسبه می‌کنند.
5. تعیین مقدار بحرانی یا :p-value با توجه به سطح معنی‌داری و درجه آزادی، مقدار بحرانی یا p-value را تعیین کنید.
6. تصمیم‌گیری :
   - اگر مقدار محاسبه شده آماره آزمون از مقدار بحرانی بزرگتر باشد یا p-value کوچکتر از سطح معنی‌داری باشد، فرض صفر رد می‌شود و نتیجه می‌گیریم که داده‌ها از یک توزیع نرمال پیروی نمی‌کنند.
   - در غیر این صورت، فرض صفر رد نمی‌شود و نمی‌توانیم با اطمینان بگوییم که داده‌ها نرمال نیستند.



نکات مهم:

- حجم نمونه :برای نمونه‌های بزرگ، حتی انحراف‌های کوچک از نرمال بودن ممکن است منجر به رد فرض صفر شود.

- توزیع‌های نزدیک به نرمال :اگر داده‌ها تقریباً نرمال باشند، ممکن است بتوان از آزمون‌های پارامتریک استفاده کرد.
- تبدیل داده‌ها :در برخی موارد، با استفاده از تبدیل‌های مناسب (مانند لگاریتم یا جذر) می‌توان داده‌ها را به توزیع نرمال نزدیک‌تر کرد.
- آزمون‌های ناپارامتریک :اگر داده‌ها به طور واضح از توزیع نرمال فاصله داشته باشند، بهتر است از آزمون‌های ناپارامتریک استفاده شود.

مثال عملی:

فرض کنید می‌خواهیم بررسی کنیم که آیا قد یک گروه از افراد به طور نرمال توزیع شده است یا خیر.

1. داده‌های قد را جمع‌آوری می‌کنیم.
2. یک هیستوگرام از داده‌ها رسم می‌کنیم تا شکل کلی توزیع را مشاهده کنیم.
3. آزمون کولموگروف-اسمیرنوف را انجام می‌دهیم.
4. اگر p-value کمتر از 0.05 باشد، نتیجه می‌گیریم که قد افراد به طور نرمال توزیع نشده است.