

Session 1&2

Regularization

Deep Learning | Zahra Amini

Telegram: @zahraamini_ai & Instagram:@zahraamini_ai & LinkedIn: @zahraamini-ai

<https://zil.ink/zahraamini>

چرا آندرفیت رخ می دهد؟

① مدل بسیار ساده: اگر مدل شما پیچیدگی کافی نداشته باشد (مثلاً تعداد لایه‌ها یا نرون‌ها کم باشد)، نمی‌تواند الگوهای پیچیده داده‌ها را بیاموزد

✓ از یک مدل پیچیده‌تر استفاده کنید، مثلاً افزایش لایه‌ها، افزایش تعداد نرون‌ها در هر لایه یا استفاده از معماری‌های پیچیده‌تر مانند ... CNN, RNN, ...

② کمبود داده: اگر داده‌های آموزشی به اندازه کافی زیاد نباشند، مدل نمی‌تواند به درستی الگوهای موجود را شناسایی کند

✓ تعداد داده‌های آموزشی را افزایش دهید. این می‌تواند از طریق جمع‌آوری داده‌های بیشتر استفاده از داده‌های مصنوعی Data Augmentation با استفاده از تکنیک‌های یادگیری انتقالی باشد Transfer Learning

۳) هایپرپارامترها: انتخاب نادرست هایپرپارامترها (مثل نرخ یادگیری، اندازه بچها و تعداد ایپاکها)

میتواند منجر به آندرفیت شدن مدل شود

✓ پارامترهای هایپری را بهینهسازی کنید. میتوانید از طریق روش‌هایی بهترین ترکیب هایپرپارامترها

Grid Search, Random Search را پیدا کنید، مانند

۴) تنظیمات ناکافی: اگر مدل به اندازه کافی آموخت نبیند (مثلاً تعداد ایپاکها کم باشد)، نمیتواند

الگوهای موجود در داده‌ها را به خوبی بیاموزد

✓ میتوانید از روش‌های اعتبارسنجی متقابل برای پیدا کردن تعداد بهینه ایپاکها استفاده کنید

cross-validation

۵) ویژگی‌های نامناسب: اگر ویژگی‌های ورودی مناسب نباشند یا اطلاعات کافی را به مدل ندهند

مدل نمیتواند به درستی پیش‌بینی کند

✓ بهبود ویژگی‌های ورودی، این میتواند شامل انتخاب ویژگی‌های بهتر، استخراج ویژگی‌های جدید

یا استفاده از تکنیک‌های کاهش ابعاد مانند PCA

چرا اورفیت رخ می دهد؟

در شبکه های عصبی، بیش برآزش زمانی رخ می دهد که مدل بر روی داده های آموزشی به خوبی عملکرد دارد اما در داده های آزمون عملکرد ضعیفی دارد. این مشکل ناشی از این است که مدل بیش از حد به جزئیات داده های آموزشی می پردازد و الگوهای خاص آن را به جای الگوهای عمومی یاد می گیرد

۱ مدل بسیار پیچیده: اگر مدل بسیار پیچیده باشد، می تواند تمامی جزئیات و نویزهای موجود در داده های آموزشی را یاد بگیرد که ممکن است در داده های جدید تکرار نشوند

۲ داده های آموزشی ناکافی: وقتی که داده های آموزشی کافی نباشد، مدل می تواند به جای یادگیری الگوهای عمومی، جزئیات خاص همان داده ها را یاد بگیرد

داده‌های آموزشی نویزی: اگر داده‌های آموزشی دارای نویز یا اطلاعات غیرمرتب زیادی باشند ③

مدل ممکن است این نویزها را نیز به عنوان الگوهای مفید بیاموزد

عدم تنوع در داده‌ها: وقتی که داده‌های آموزشی تنوع کافی نداشته باشند، مدل به ④

الگوهای خاص و منحصر به فرد آن داده‌ها وابسته می‌شود

تعداد زیاد ایپاک‌ها: اگر مدل برای تعداد زیادی ایپاک آموزش داده شود، ممکن است بیش از حد ⑤

به داده‌های آموزشی وابسته شود

راه حل هایی برای جلوگیری از اورفیت

- ✓ مدل را ساده تر کنید تا فقط الگوهای عمومی و مهم را یاد بگیرد (کاهش تعداد لایه ها یا نرون ها باشد)
- ✓ افزایش داده های آموزشی: با جمع آوری داده های بیشتر، مدل می تواند الگوهای عمومی تر و کمتر وابسته به جزئیات خاص یاد بگیرد. همچنین می توانید از تکنیک های Data Augmentation برای تولید داده های مصنوعی استفاده کنید
- ✓ از تکنیک های تنظیم گر استفاده کنید تا مدل نتواند بیش از حد به داده های آموزشی وابسته شود L1, L2 Regularization
- ✓ استفاده از روش Dropout ، در این روش در هر مرحله آموزش، تعدادی از نuron ها به صورت تصادفی حذف می شوند. به عبارت دیگر، در هر تکرار از فرآیند آموزش به صورت رندوم نuron های مختلفی موقتاً غیرفعال می شوند

✓ توقف زودهنگام یا Early Stopping

این روش باعث می‌شود که آموزش مدل زمانی که عملکرد روی داده‌های اعتبارسنجی شروع به کاهش می‌کند، متوقف شود

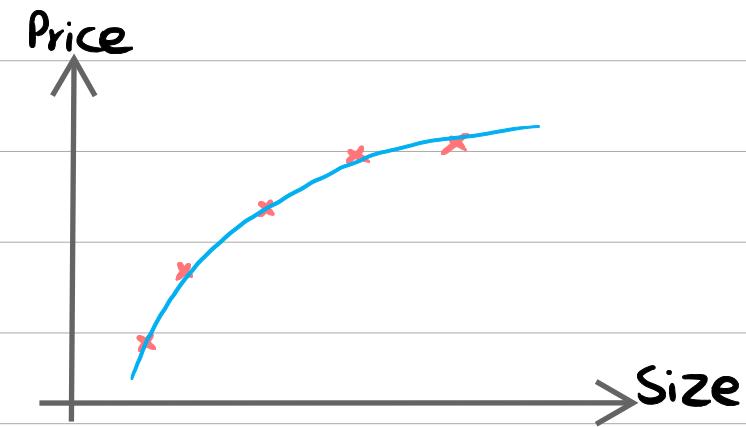
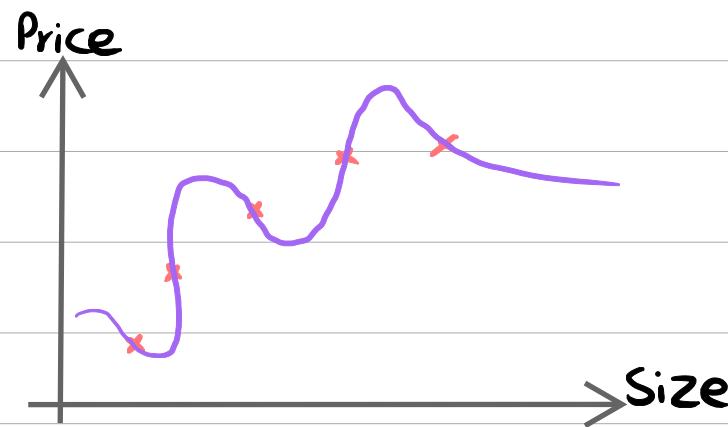
✓ تقسیم تصادفی داده‌ها یکی از اصول مهم در یادگیری ماشین است که به مدل کمک می‌کند تا الگوهای عمومی‌تر را یاد بگیرد و از اورفیت شدن جلوگیری کند. عدم انجام این کار می‌تواند به نتایج نادرست و عملکرد ضعیف مدل در داده‌های واقعی منجر شود

✓ تعداد اپوک‌ها را با توجه به پیچیدگی مدل و حجم داده‌های آموزشی تنظیم کنید. می‌توانید از روش‌های اعتبارسنجی متقابل برای پیدا کردن تعداد بهینه اپوک‌ها استفاده کنید

Regularization:

دا توانه از overfit , underfit جلوگیری کند. چه طور؟

که می کند ضرایب را کوچک یا حدف کنند.

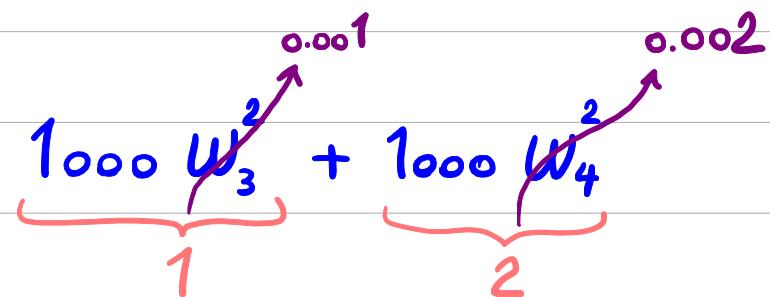


$$W_1x + W_2x^2 + W_3x^3 + W_4x^4 + b$$

$$W_1x + W_2x^2 + \underbrace{W_3x^3}_{\approx 0} + \underbrace{W_4x^4}_{\approx 0} + b$$

Make W_3, W_4 really small (≈ 0)

$$\min \left[\frac{1}{2m} \sum_{i=1}^m (f_{w,b}(x^{(i)}) - y^{(i)})^2 \right]$$



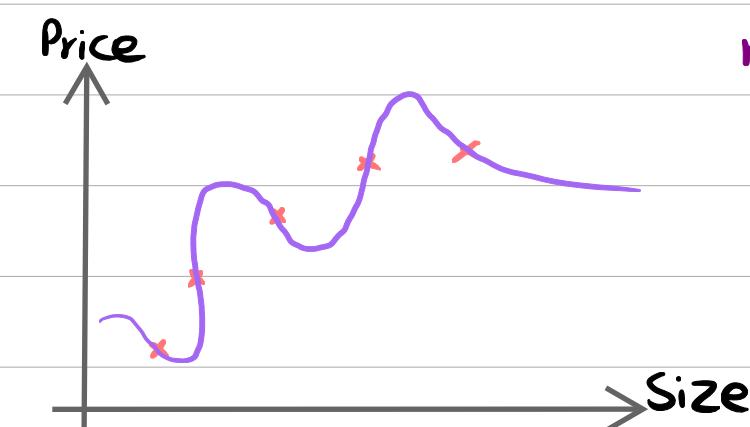
x_1	x_2	x_3	x_4	x_5	\dots	x_{100}	y
Size	bedroom	floors	age	avg income	...	distance coffee shop	Price

$w_1, w_2, w_3, \dots, w_{100}, b$

$n=100$

$$J(w, b) = \frac{1}{2m} \sum_{i=1}^m (f_{w,b}(x^{(i)}) - y^{(i)})^2 + \frac{\lambda}{2m} \sum_{j=1}^n w_j^2 + \frac{\lambda}{2m} b^2$$

λ : Lambda Regularization Parameter



$$\min J(w, b) = \min \left[\frac{1}{2m} \sum_{i=1}^m (f_{w,b}(x^{(i)}) - y^{(i)})^2 + \frac{\lambda}{2m} \sum_{j=1}^n w_j^2 \right]$$

$w, b = ?$

Gradient Descent:

repeat

{

$$w_j = w_j - \alpha \frac{\partial}{\partial w_j} J(w, b) \rightarrow = \frac{1}{m} \sum_{i=1}^m (f_{w,b}(x^{(i)}) - y^{(i)}) x_j^{(i)} + \frac{\lambda}{m} w_j$$

Regularizer term

$$b = b - \alpha \frac{\partial}{\partial b} J(w, b) \rightarrow = \frac{1}{m} \sum_{i=1}^m (f_{w,b}(x^{(i)}) - y^{(i)})$$

}

$$\frac{\partial}{\partial w_j} \left[\frac{1}{2m} \sum_{i=1}^m \underbrace{(f_{w,b}(x^{(i)}) - y^{(i)})^2}_{w \cdot x^{(i)} + b} + \frac{\lambda}{2m} \sum_{j=1}^n w_j^2 \right] = \frac{1}{2m} \sum_{i=1}^m \left[(wx^{(i)} + b - y^{(i)}) \times 2 \times x_j^{(i)} \right] + \sum_{j=1}^n \frac{\lambda}{2m} \times 2 w_j$$

$$= \frac{1}{m} \sum_{i=1}^m \underbrace{\left[(w \cdot x^{(i)} + b - y^{(i)}) x_j^{(i)} \right]}_{f_{w,b}(x)} + \sum_{i=1}^m \frac{\lambda}{m} w_j = \frac{1}{m} \sum_{i=1}^m \left[(f_{w,b}(x^{(i)}) - y^{(i)}) x_j^{(i)} \right] + \frac{\lambda}{m} w_j$$

$$w_j = w_j - \alpha \frac{1}{m} \sum_{i=1}^m (f_{w,b}(x^{(i)}) - y^{(i)}) x_j^{(i)} + \frac{\lambda}{m} w_j$$

$$w_j = w_j - \underbrace{\alpha \frac{\lambda}{m} w_j}_{w_j(1-\alpha \frac{\lambda}{m})} - \underbrace{\alpha \frac{1}{m} \sum_{i=1}^m (f_{w,b}(x^{(i)}) - y^{(i)}) x_j^{(i)}}_{\text{usual update}}$$

$$\hookrightarrow \alpha = 0.01, \lambda = 1 \implies \alpha \frac{\lambda}{m} = 0.01 \times \frac{1}{50} = 0.0002$$

0.9998

$$w_j \cancel{(1-0.0002)}$$

Regularization

Ridge

$$L2 = \sum_{j=0}^n w_j^2$$

alpha $\rightarrow \alpha$

Lasso

$$L1 = \sum_{j=0}^n |w_j|$$

Elastic Net

L1, L2

Ridge \rightarrow معمولی سرایب کوچک ہی شوند.

$\alpha = 0 \rightarrow$ Ridge X

$$\text{Loss} = \sum_{i=1}^m (f_{w,b}(x) - y^{(i)})^2 + \lambda \sum_{j=0}^n w_j^2$$

$\alpha = \infty \rightarrow$ معمولی سرایب کوچک و نزدیک بہ صفر ہی شوند.

ضمیراییب را کوچک کرند و بضرفی از ضمیراییب را به ۵ کارساند.

$$\text{Loss} = \sum_{i=1}^m (f_{w,b}(x) - y^{(i)})^2 + \lambda \sum_{j=0}^n |w_j|$$

$\alpha = 0 \rightarrow$ هیچ وزنی حذف نباشد.

$\alpha = \infty \rightarrow$ تمام وزنها را حذف کنند.

Elastic Net:

↳ L1, L2

lasso , Ridge ترکیبی است از

$$\text{Loss} = \frac{1}{2m} \sum_{i=1}^m (f_{w,b}(x) - y_i)^2 + (1-\lambda) \times \frac{\alpha}{2} \times \underbrace{\sum_{i=1}^n w_i^2}_{L2} + \underbrace{\lambda \alpha |w_i|}_{L1}$$

L1_rate = 1 \rightarrow L1 \rightarrow lasso

L1_rate = 0 \rightarrow L2 \rightarrow Ridge