

آموزش شبکه‌های عصبی

تعمیم‌دهی (Generalization)

جلسه دوم و سوم

مهرسا یزدانی

تابستان ۱۴۰۴

محتوا

تعییندهی

۳ سایر روش‌ها

Dropout ۳

۳ مدل‌های گروهی

۲ منظم سازی

۱ مرور

Batch Normalization

۴ تعریف

مجموع داده‌ی نامتوابن

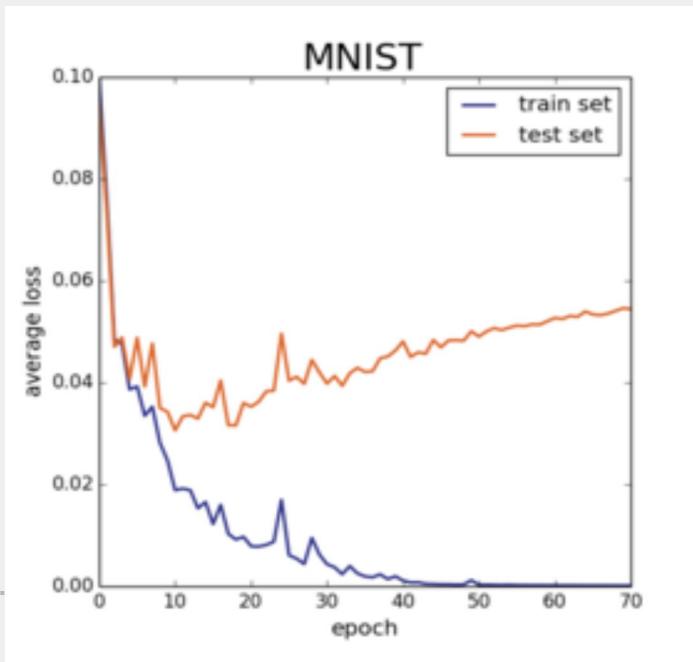
۵ روش‌های مدیریت داده نامتوابن

۴ تعریف

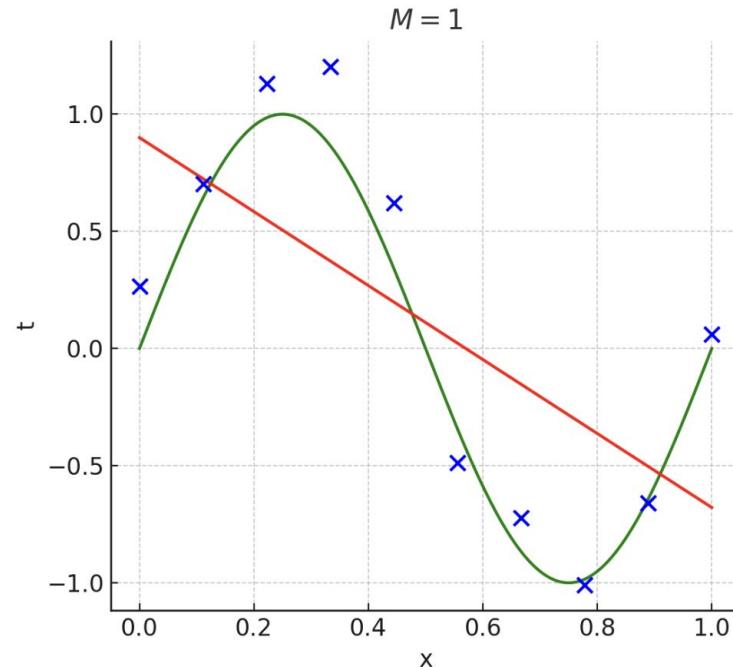
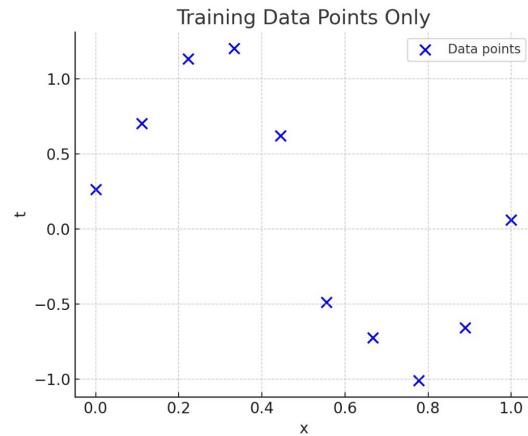
جمع‌بندی

نمایمدهی چیست؟

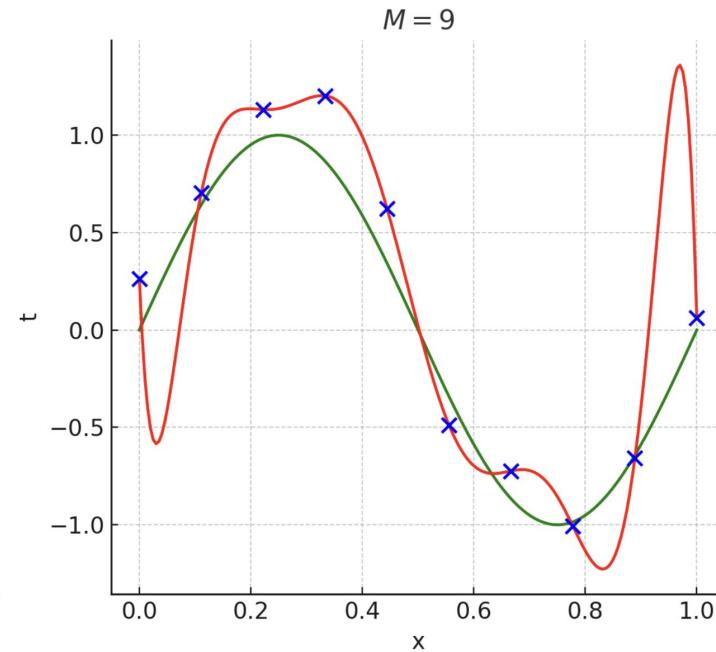
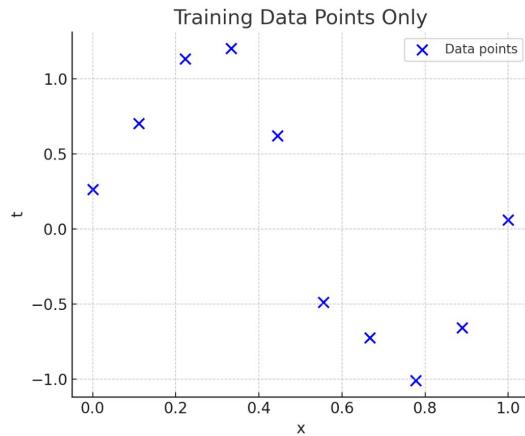
می خوایم فاصله‌ی بین خطای آموزش و تست کم شود. یعنی مدل هم روی داده‌ی آموزش هم روی داده‌ی تست خوب عمل کند.



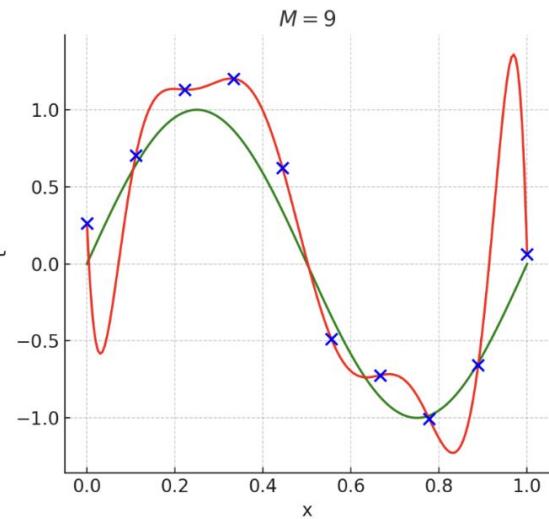
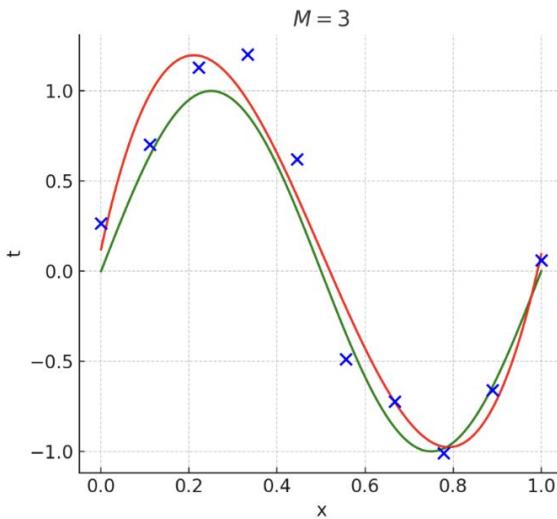
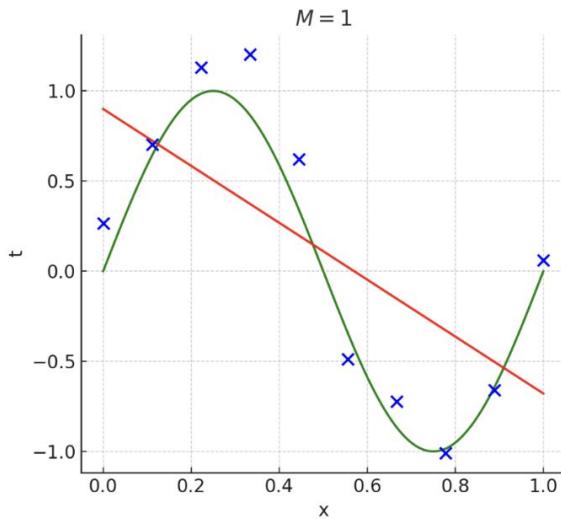
Underfitting



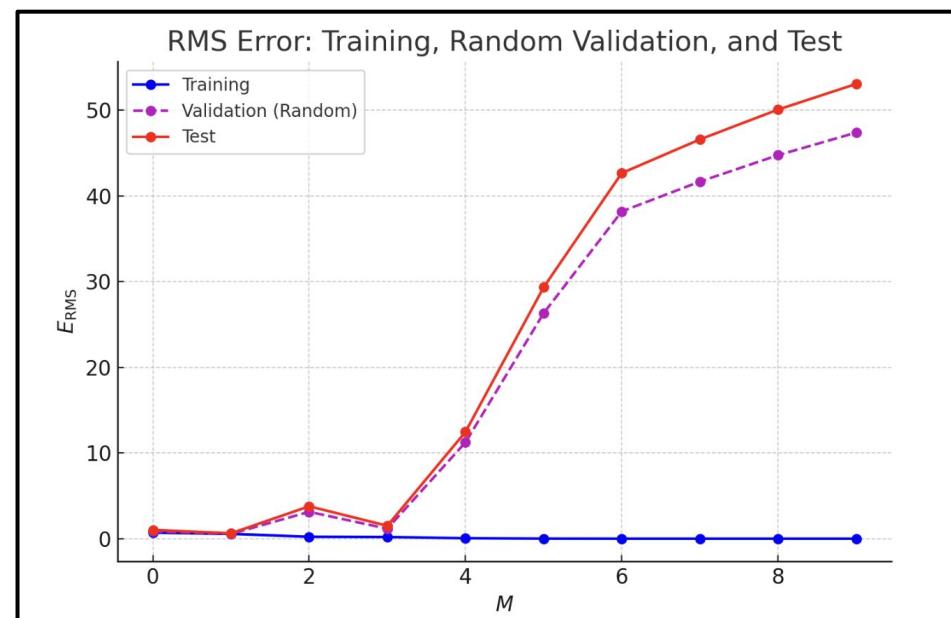
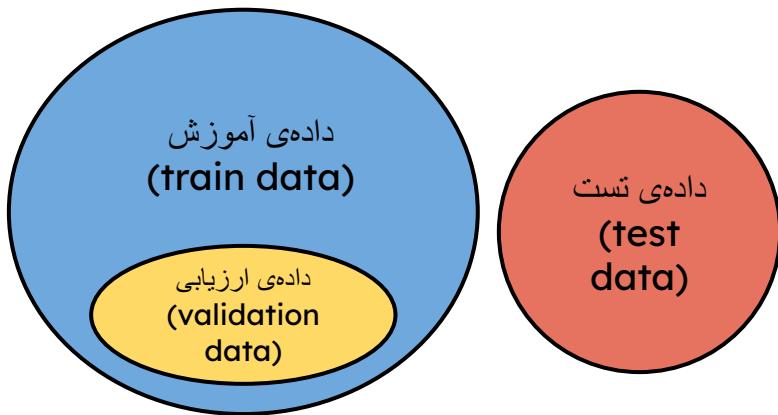
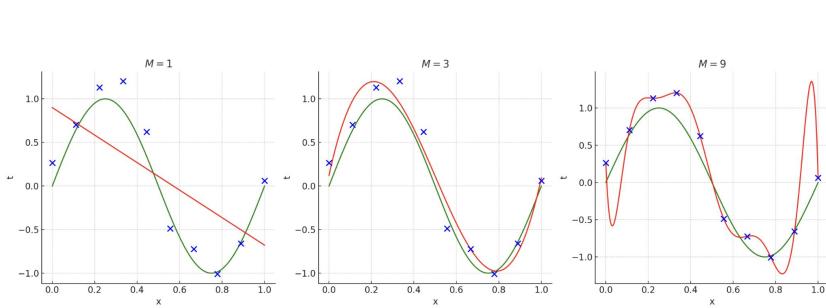
Overfitting



انتخاب مدل درست



انتخاب مدل درست





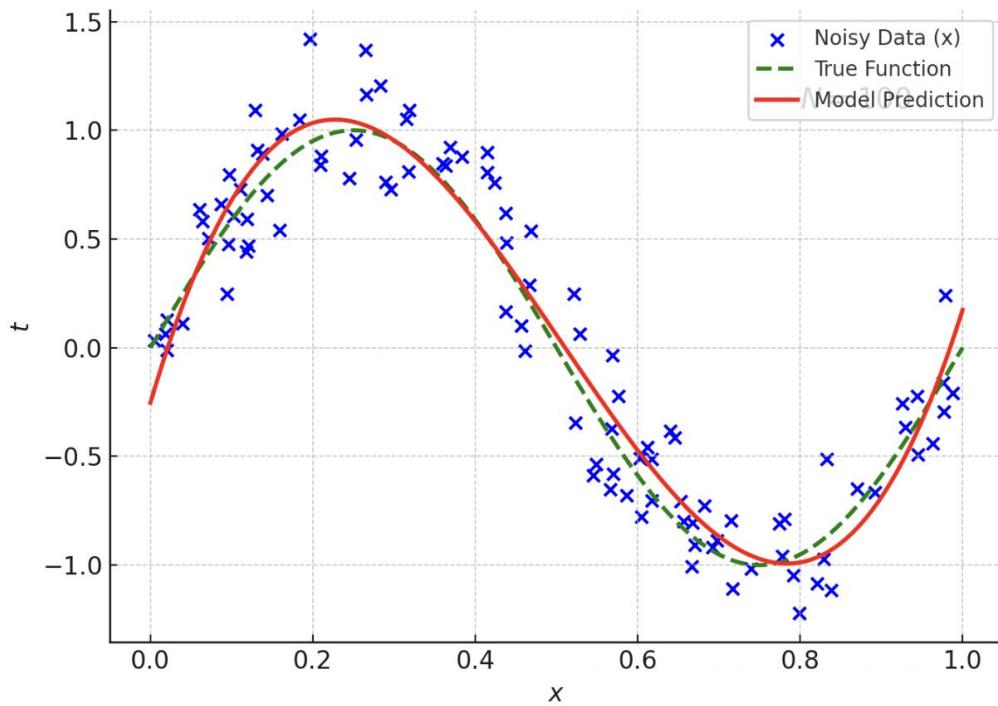
روش‌های افزایش تعمیمدهی

افزایش داده

(Data Augmentation)



افزایش داده



افزایش داده

داده‌های آموزشی در بسیاری از موقع کم هستند.
با اعمال تغییرات مختلف روی نمونه‌ها، می‌توان آن‌ها را «گسترش» داد و نمونه‌های برچسبدار مصنوعی تولید کرد.



CocaColaZero1_1.png



CocaColaZero1_2.png



CocaColaZero1_3.png



CocaColaZero1_4.png



CocaColaZero1_5.png



CocaColaZero1_6.png



CocaColaZero1_7.png



CocaColaZero1_8.png

چه زمان نباید از افزایش داده استفاده کرد؟

چه زمان نباید از افزایش داده استفاده کرد؟

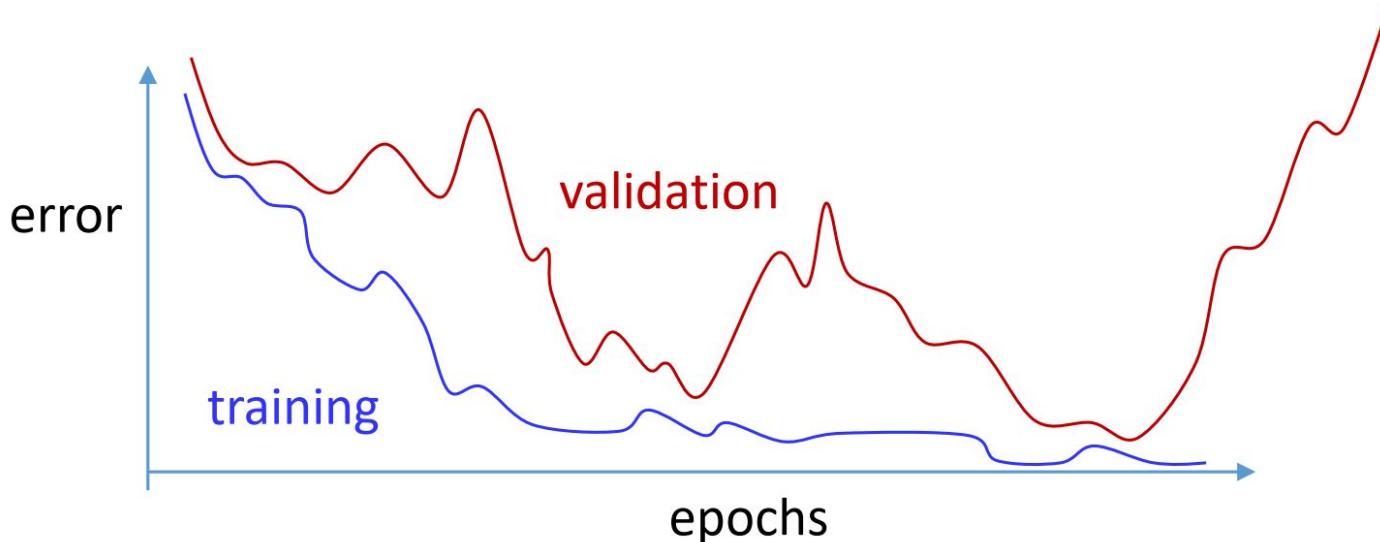
- از قبل تعداد زیادی داده‌ی واقعی وجود دارد
- زمانی که با داده‌های حساس سر و کار داریم. (مانند زمانی که تغییرات کوچک برچست داده را عوض می‌کند)
- زمانی که تغییرات داده می‌تواند معنای داده را تغییر دهد. (مانند چرخاندن عکس‌های پزشکی)
- در زمان استفاده از داده‌ی ارزیابی یا داده‌ی تست



۲ توقف زودهنگام
(Early Stopping)

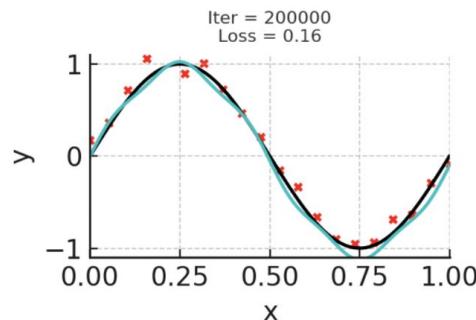
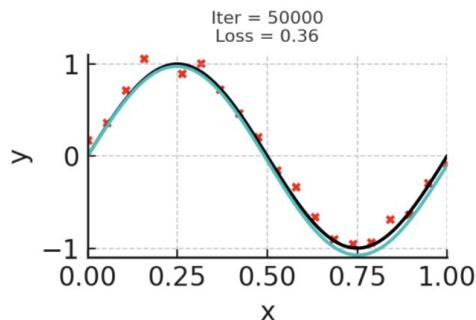
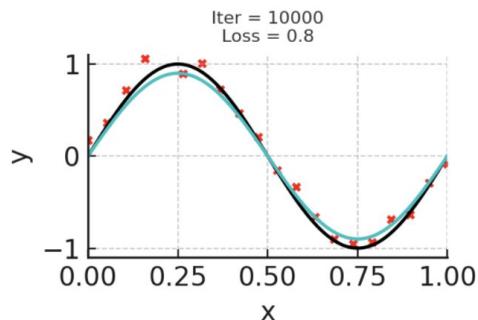
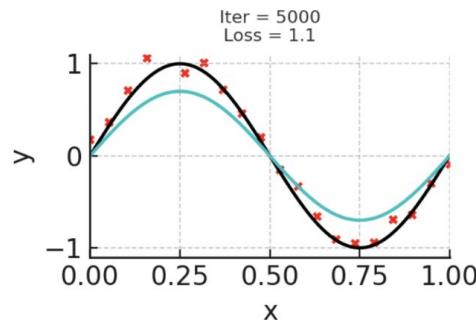
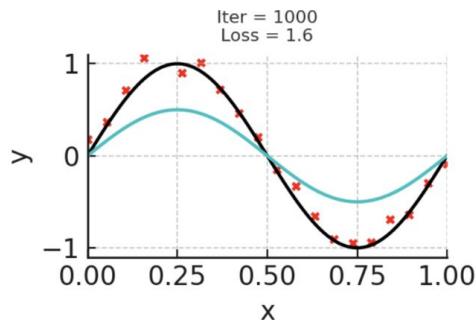
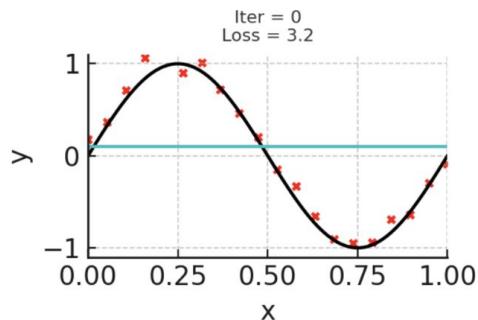
توقف زودهنگام

- ادامه دادن آموزش می‌تواند منجر به Overfit شدن به داده‌های آموزش شود.
- می‌توان عملکرد مدل را بر روی داده‌ی ارزیابی دنبال کرد.
- می‌توان یکی از چندین معیار توقف زودهنگام را برای پایان دادن به آموزش، زمانی که عملکرد بر روی داده‌ی ارزیابی به طور قابل توجهی کاهش می‌یابد، اعمال کرد.



توقف زودهنگام

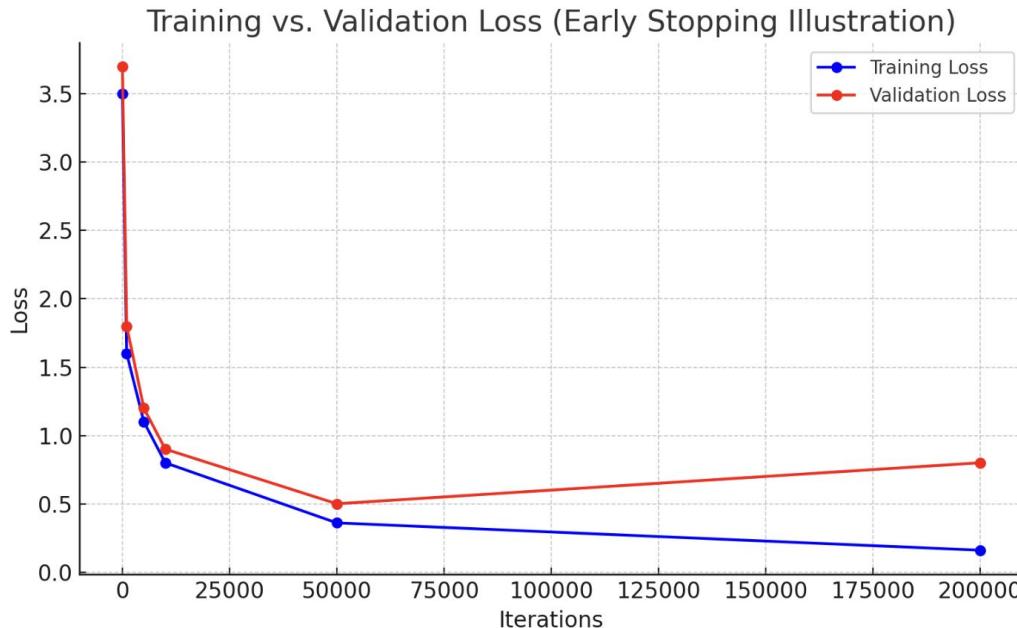
Early Stopping Example (Training Progress)



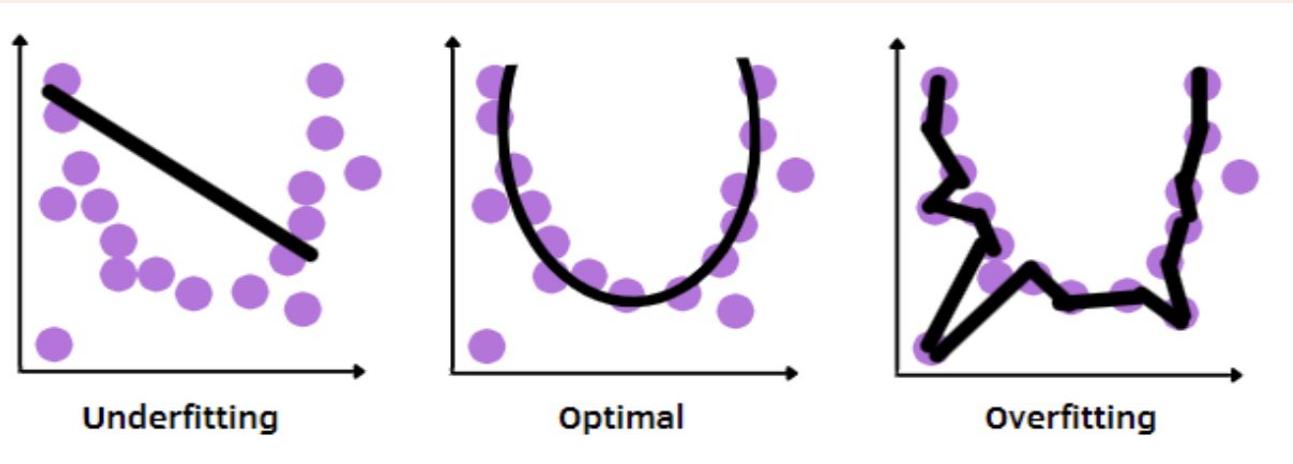
مثال از زمان آموزش

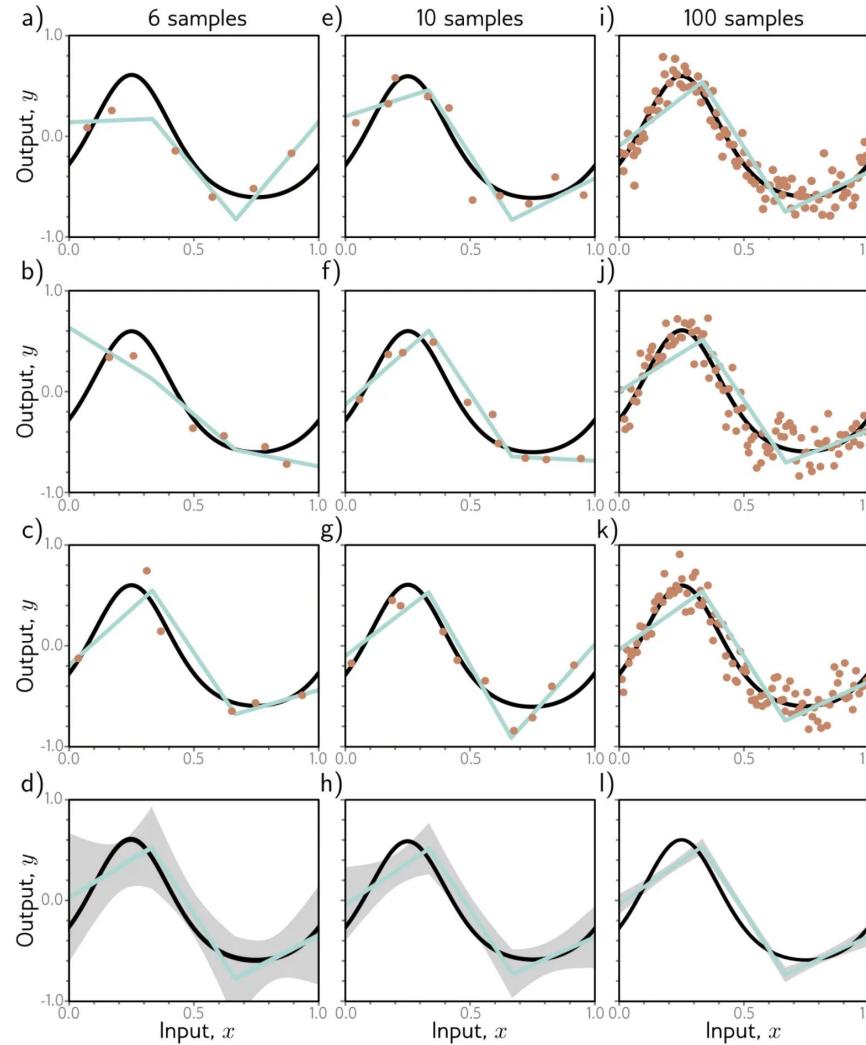
توقف زودهنگام

نمودار خطای آموزش و ارزیابی



منظم سازی (Regularization)



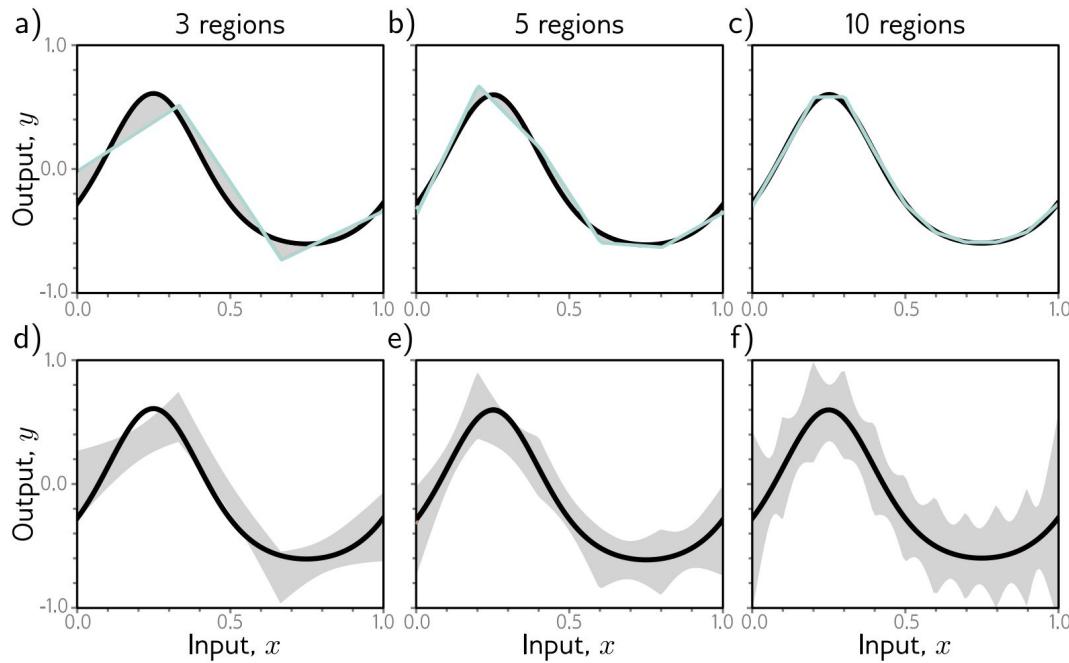


واریانس (Variance)

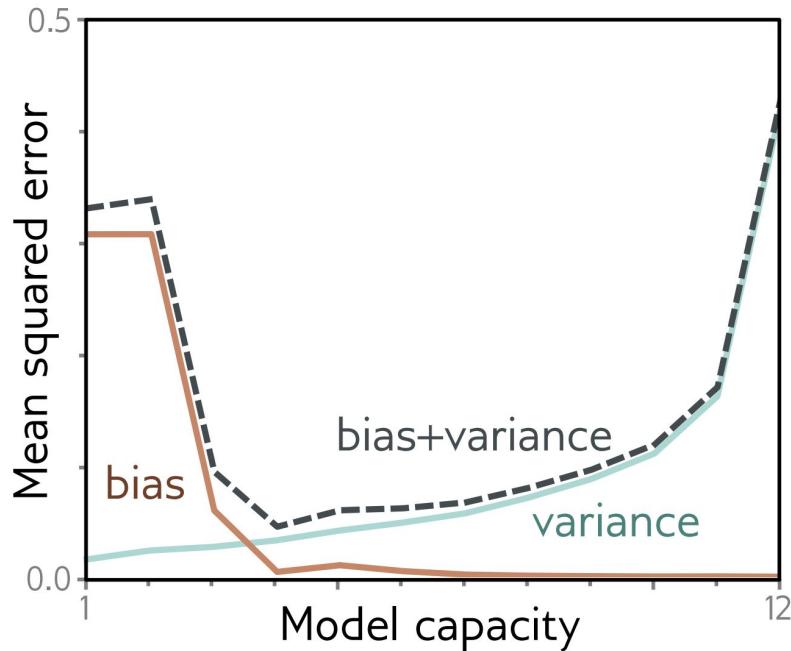
واریانس مدل آموزش داده شده بر روی مجموعه
نمونه‌های مختلف

بایاس (Bias)

بایاس در مقایسه با واریانس



Bias & Variance Trade off



هر چه بایاس کمتر می‌شود، واریانس بیشتر می‌شود.

- با واریانس بیشتر چه اتفاقی برای Generalization می‌افتد؟

Regularization

- چگونه می‌توانیم فاصله‌ی خطای آموزش و تست را کم کنیم؟
- اضافه کردن یک جمله به تابع هزینه برای افزودن inductive bias به آن.

$$J(W) = \sum_{n=1}^N \text{loss}\left(y^{(n)}, f(y^{(n)}, W)\right) + \lambda R(W)$$

L2 Regularization

$$J(W) = \sum_{n=1}^N \text{loss}\left(y^{(n)}, f(y^{(n)}, W)\right) + \lambda R(W)$$

$$J(W) = L(W) + \lambda R(W)$$

$$J(W) = L(W) + \lambda \sum_{l=1}^L \|W^{[l]}\|^2$$

$$W^{[l]} \leftarrow W^{[l]} - \alpha \nabla_{W^{[l]}} L(W) - 2\lambda W^{[l]}$$

$$W^{[l]} \leftarrow W^{[l]} (1 - 2\lambda) - \alpha \nabla_{W^{[l]}} L(W)$$

Weight decay

- با اضافه کردن Regularization به یک نمودار نرمتر هزینه بر حسب epoch خواهیم رسید.

ضریب λ

$$J(W) = L(W) + \lambda R(W)$$

$$J(W) = L(W) + \lambda \sum_{l=1}^L \|W^{[l]}\|^2$$

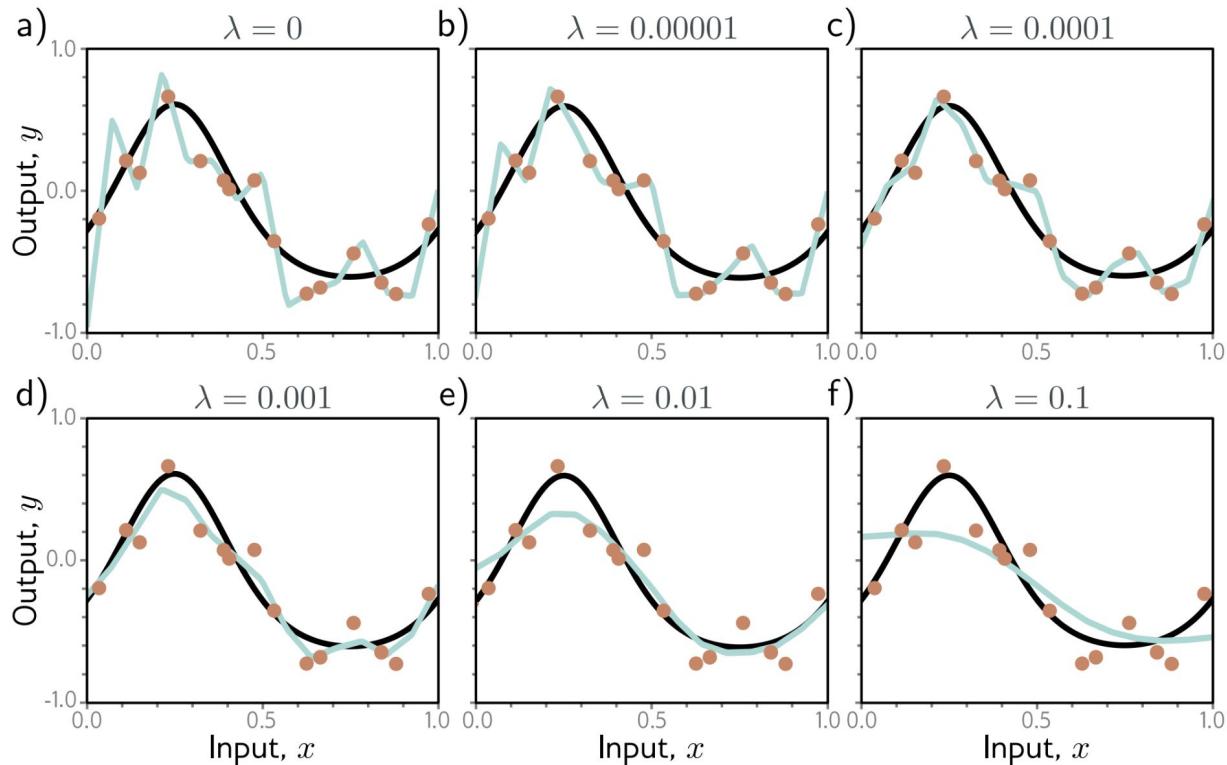
$$W^{[l]} \leftarrow W^{[l]} - \alpha \nabla_{W^{[l]}} L(W) - 2\lambda W^{[l]}$$

$$W^{[l]} \leftarrow W^{[l]}(1 - 2\lambda) - \alpha \nabla_{W^{[l]}} L(W)$$

Weight decay

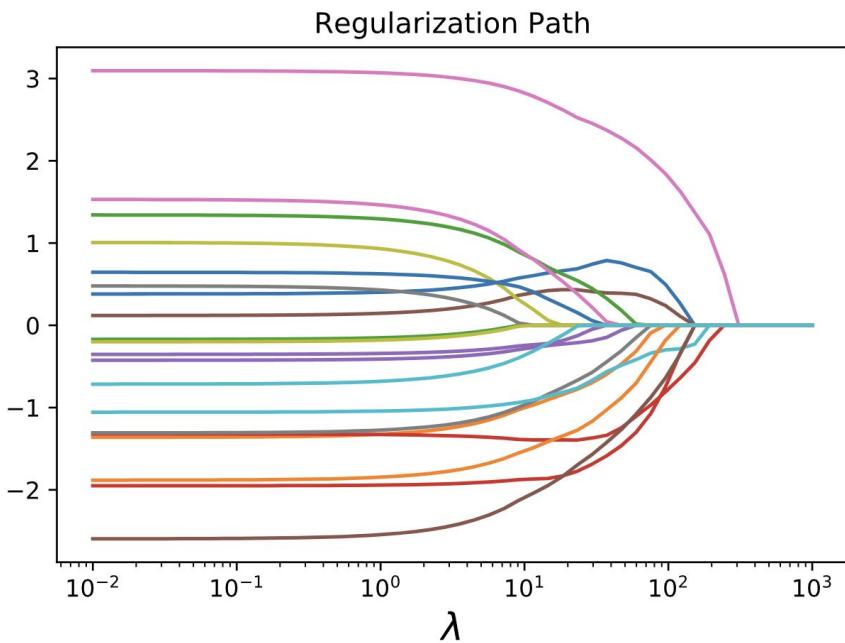
- هرچه این ضریب بزرگتر باشد یعنی برایمان مهمتر است که وزن‌ها کاهش پیدا کنند.

Regularization اثر



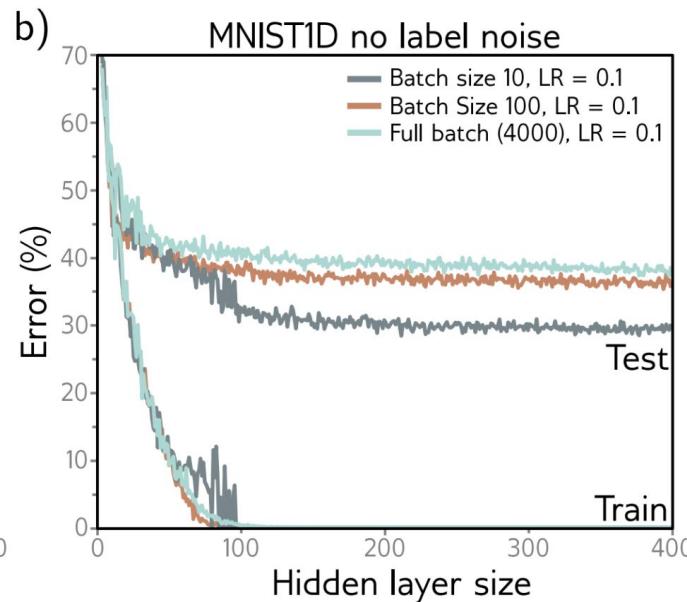
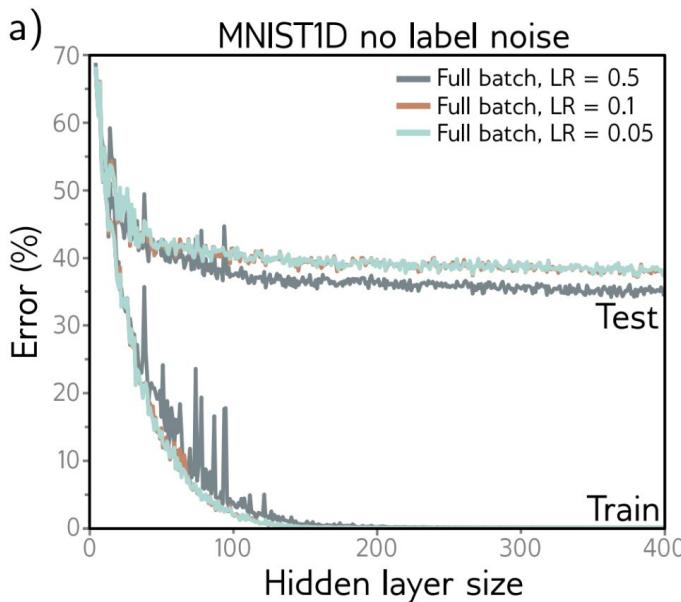
L1 Regularization

$$J(W) = L(W) + \lambda \sum_{l=1}^L \|W^{[l]}\|$$



سایر حالت‌های Regularization

نحوه‌ی Optimization می‌تواند اثر منظم‌سازی داشته باشد.

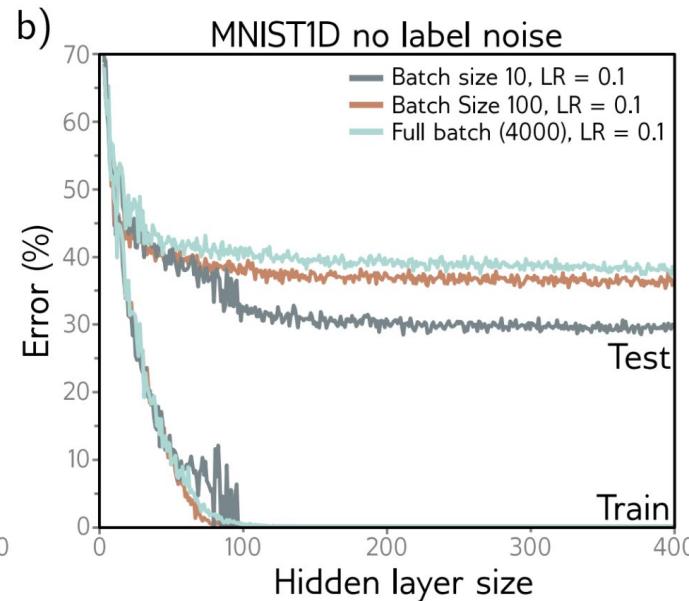
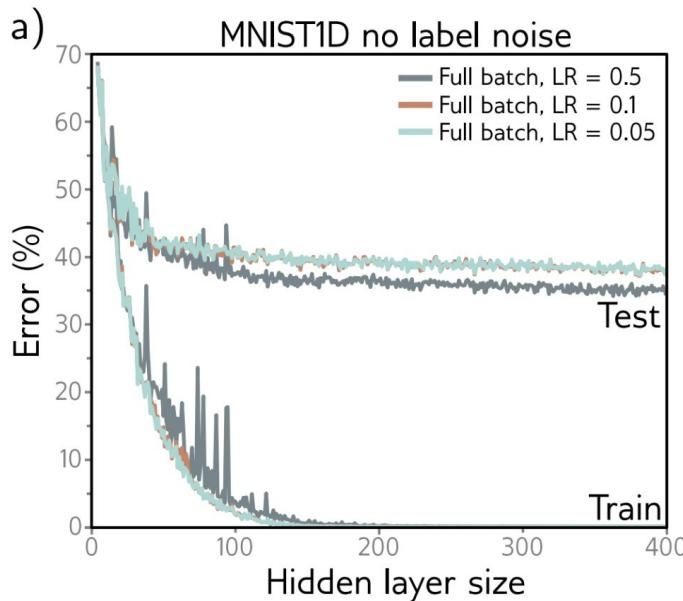


سایر حالت‌های Regularization

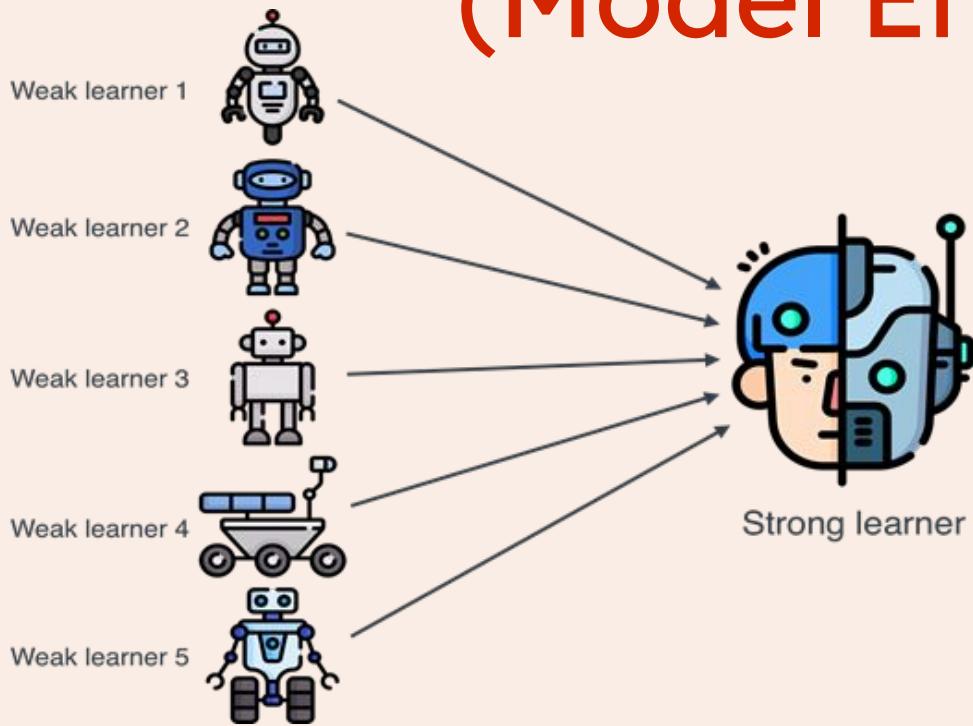
خود Early stopping چطور؟

نحوه‌ی Optimization می‌تواند اثر منظم‌سازی

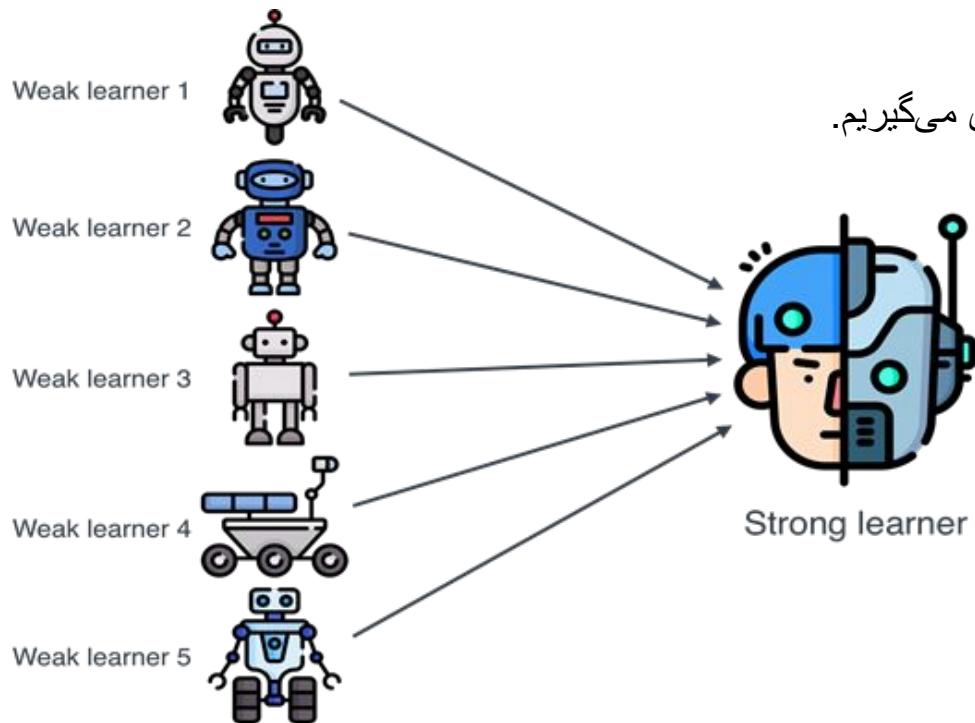
داشته باشد.



مدل‌های گروهی (Model Ensembles)



مدل‌های گروهی



تعدادی مدل مستقل از هم داشته باشیم.

در زمان تست از خروجی آن‌ها میانگین می‌گیریم.

-
-

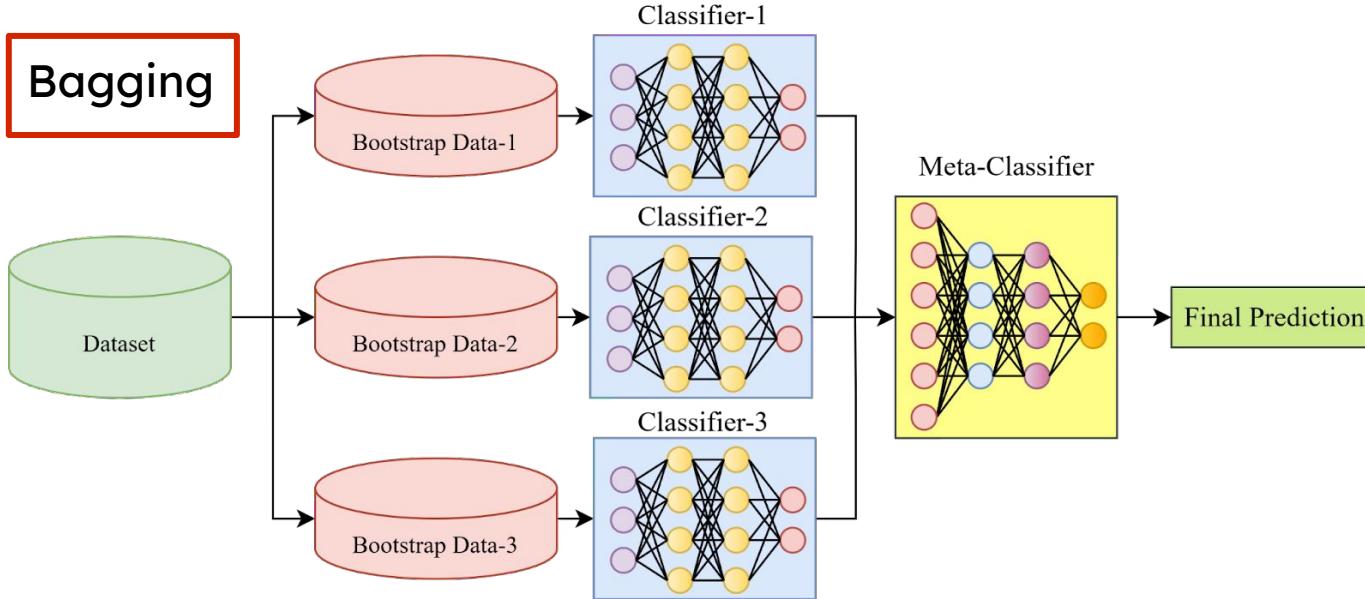
مدل‌های گروهی

کاندیداهای مختلف:

- ○ مدل‌های گوناگونی آموزش دهیم. (یا خود مدل متفاوت باشه یا داده‌ی آموزش داده شده.)
- مدل‌های شبیه به هم استفاده کنیم ولی نقطه شروع متفاوتی در نظر بگیریم.
- بهترین مدل‌های زمان cross-validation را استفاده کنیم. (هر کدام به نوعی خوب عمل کرده‌اند.)
-

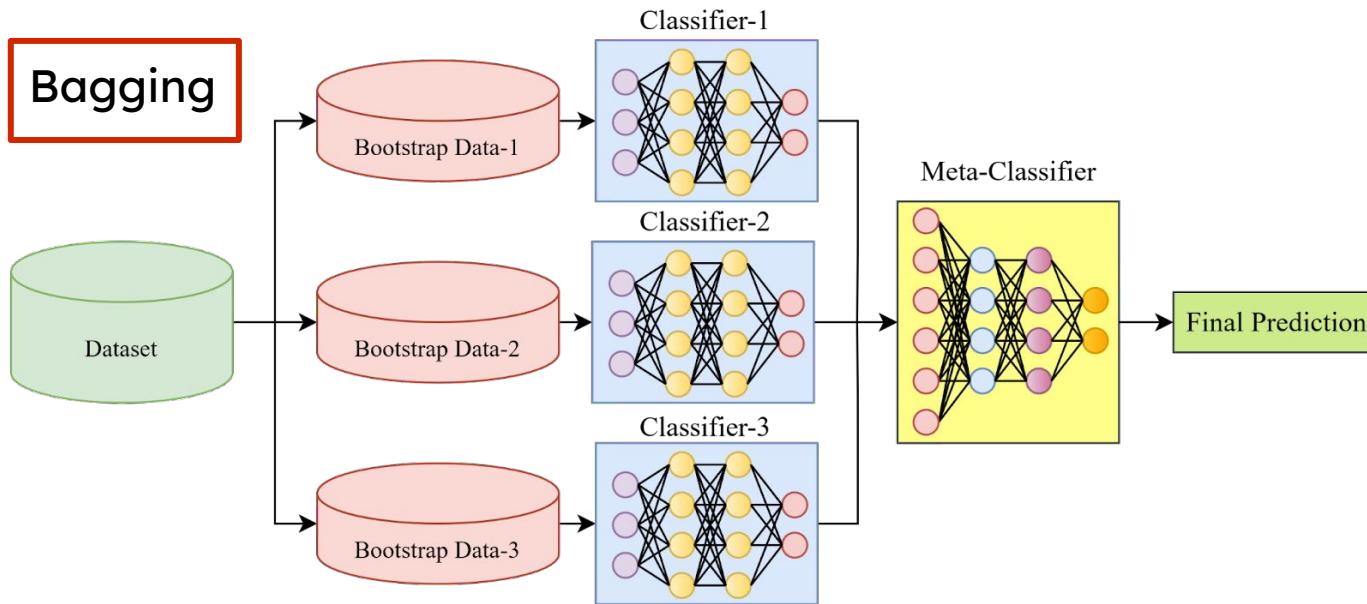
چرا ensemble کردن باعث افزایش Generalization می شود؟

میانگین گیری باعث کاهش واریانس می شود.



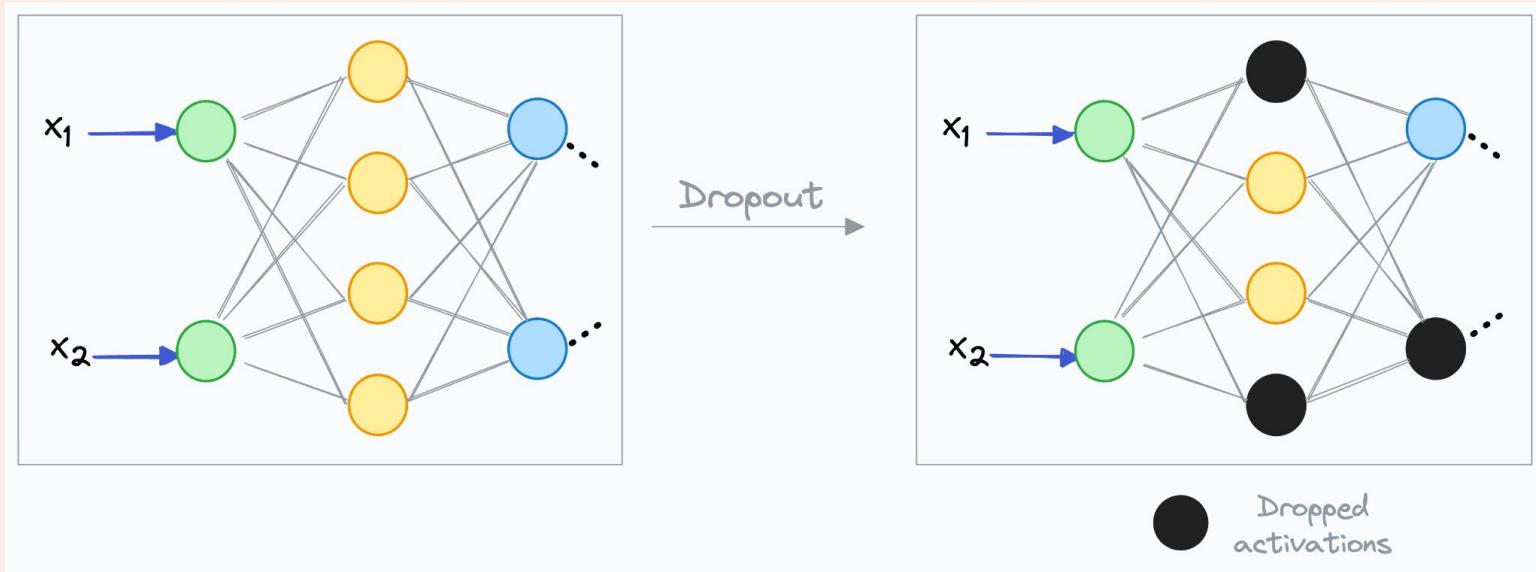
چرا ensemble کردن باعث افزایش Generalization می شود؟

تست



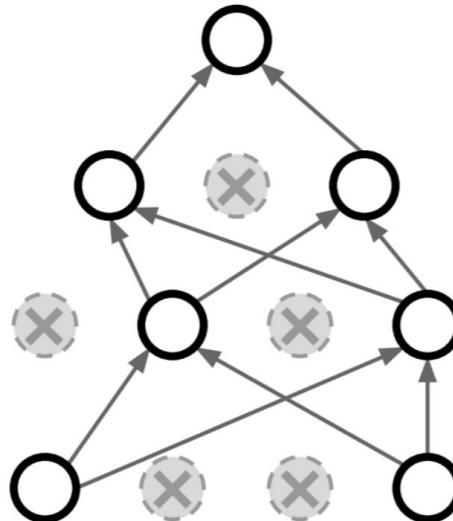
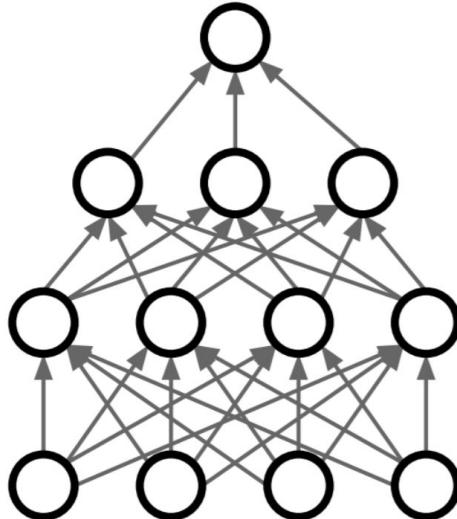
Dropout

5



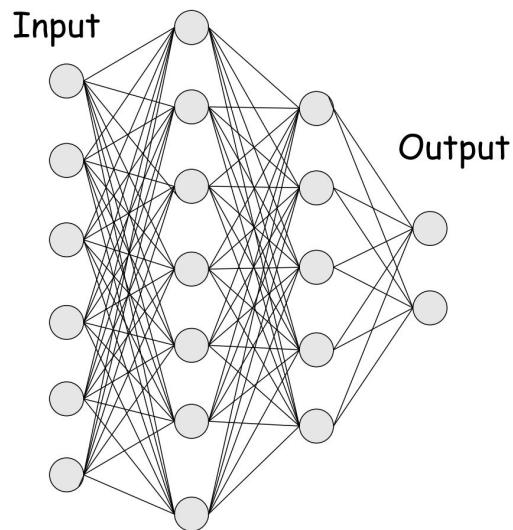
چیست؟ Dropout

- در هر forward pass به صورت رندوم تعدادی از نورون‌ها را صفر می‌کنیم.
- احتمال حذف کردن یک نورون یک هایپرپارامتر است و معمولاً 0.5 در نظر گرفته می‌شود.



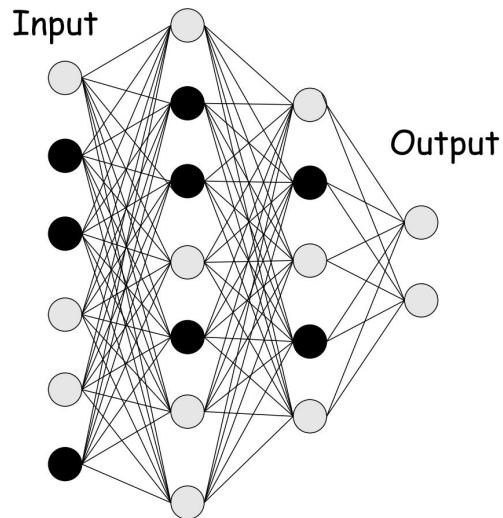
Dropout چیست؟

- در زمان training: برای هر ورودی، خاموش کن هر نورون را با احتمال $\alpha-1$

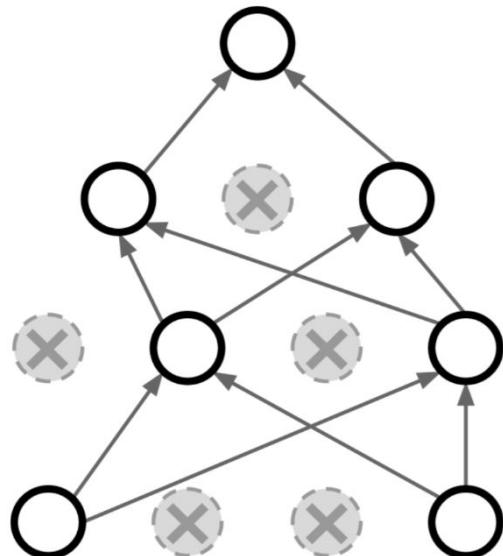


Dropout چیست؟

- در زمان training: برای هر ورودی، خاموش کن هر نورون را با احتمال $\alpha-1$



چگونه می‌تواند ایده‌ی خوبی Dropout باشد؟

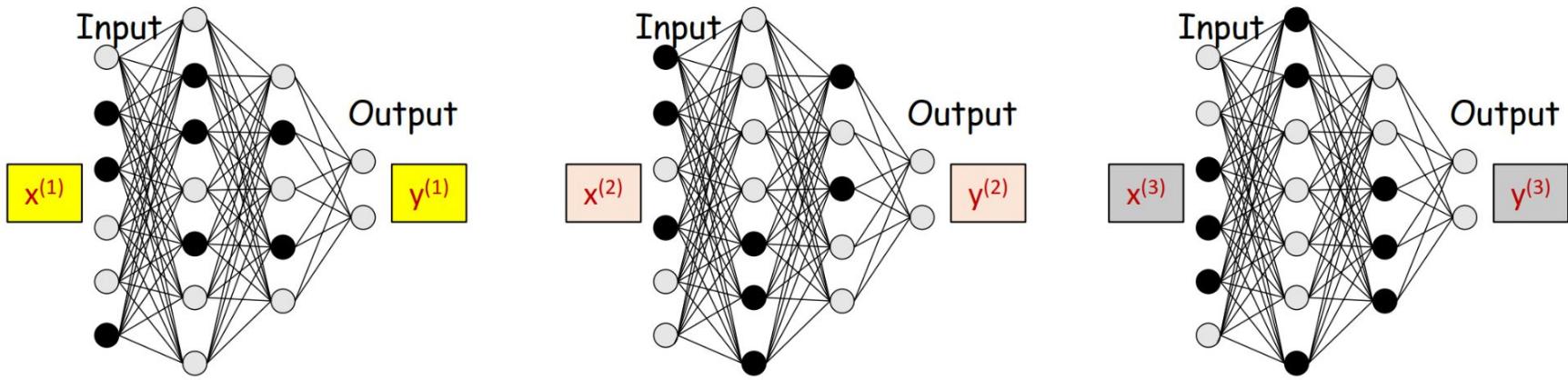


Forces the network to have a redundant representation;
Prevents co-adaptation of features

[Hinton et al., 2012]

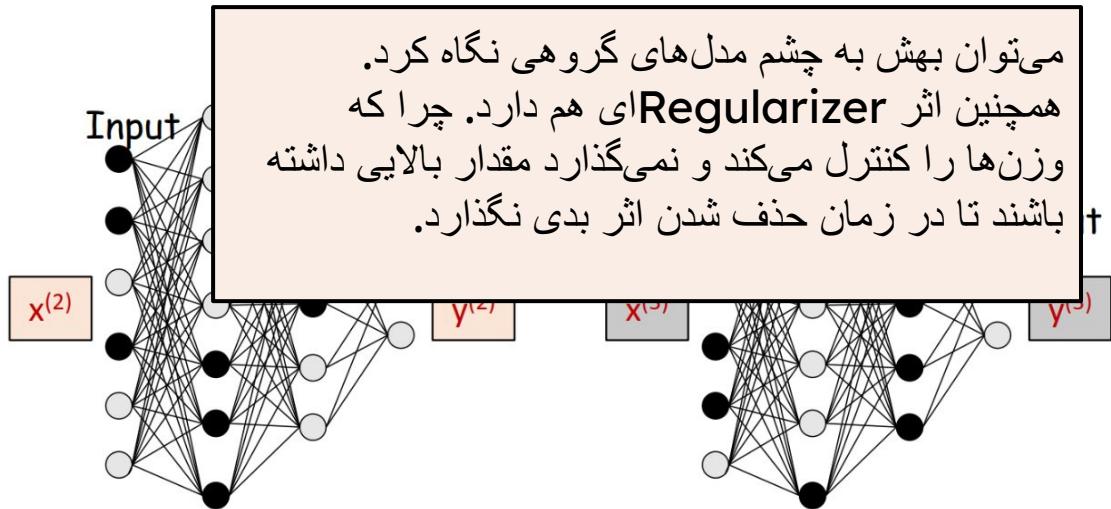
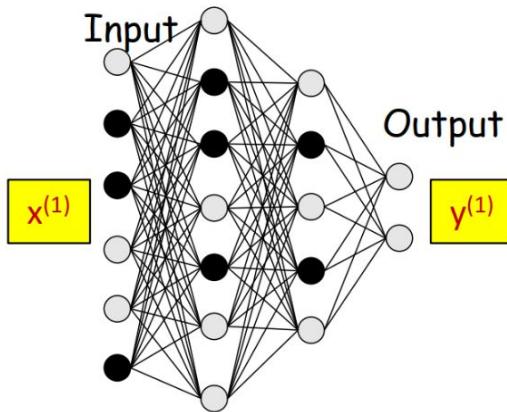


چگونه می‌تواند ایدهٔ خوبی Dropout باشد؟



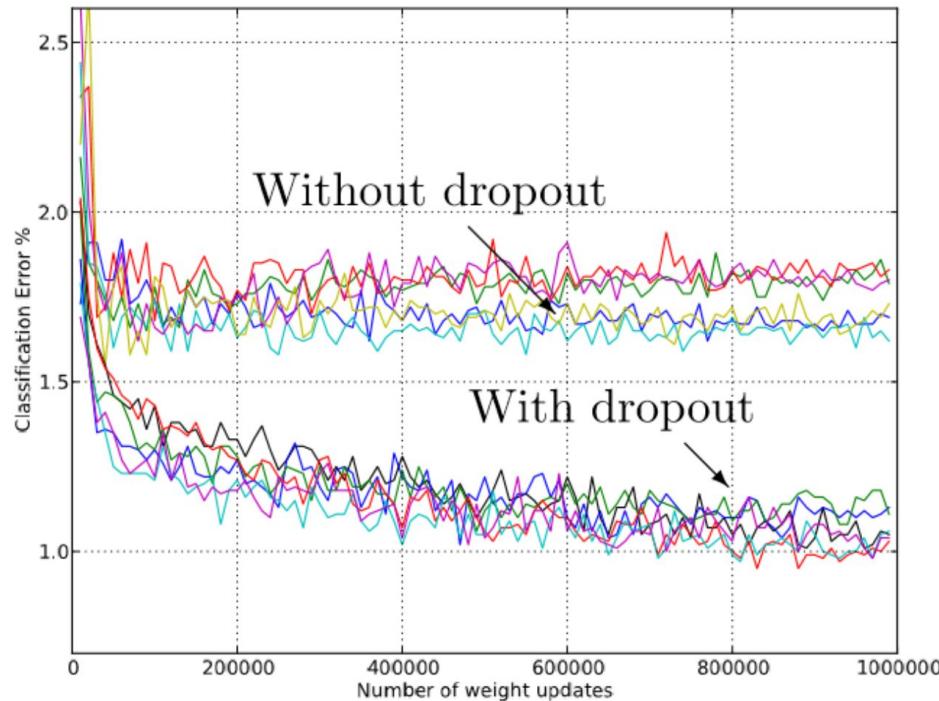
The pattern of dropped nodes
changes for each input
i.e. in every pass through the net

چگونه می‌تواند ایدهٔ خوبی Dropout باشد؟



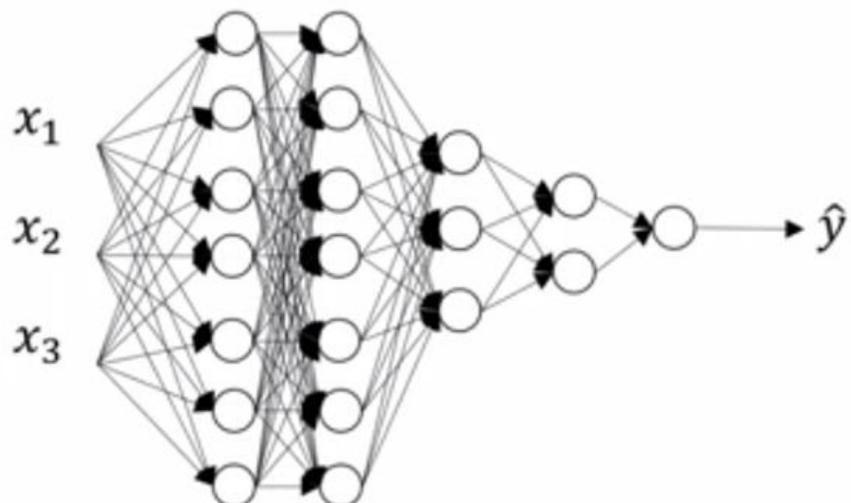
The pattern of dropped nodes
changes for each input
i.e. in every pass through the net

چگونه می‌تواند ایدهٔ خوبی Dropout باشد؟



انواع Dropout

- احتمال حذف کردن یک نورون در لایه‌های مختلف متفاوت باشد.



اگر Dropout نتیجه را بهتر نکرد؟

اگر Dropout نتیجه را بهتر نکرد؟

شبکه ضعیف و کوچک است و با افزایش تعداد پارامترها و بعد اعمال Dropout نتیجه بهتر خواهد شد.

سایر روش‌های

افزایش تعمیمدهی

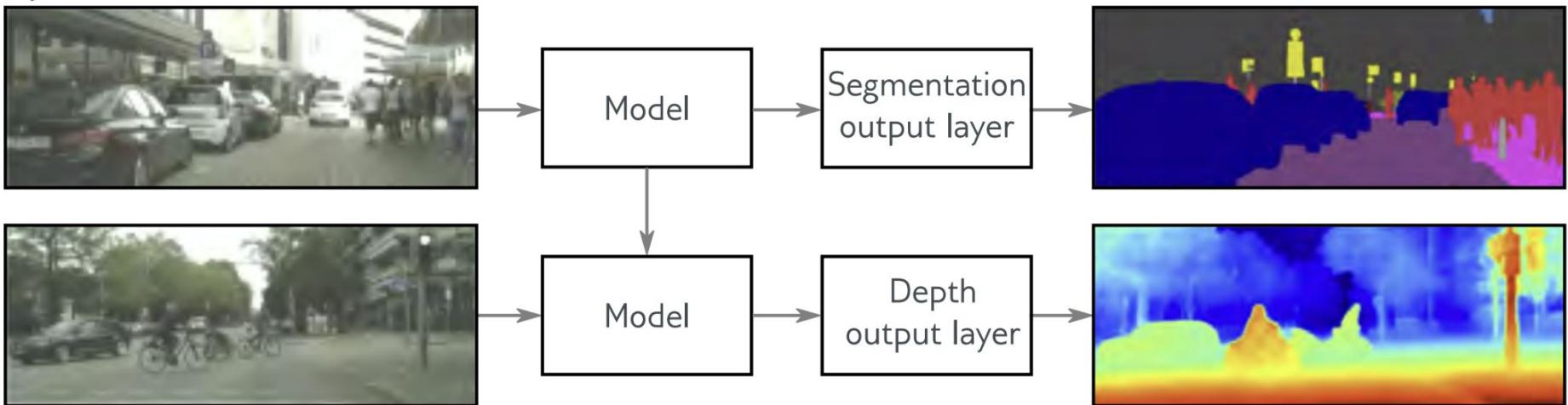
اضافه کردن نویز

- بر روی ورودی (داده)
- بر روی وزن‌های شبکه
- بر روی برچسب‌ها

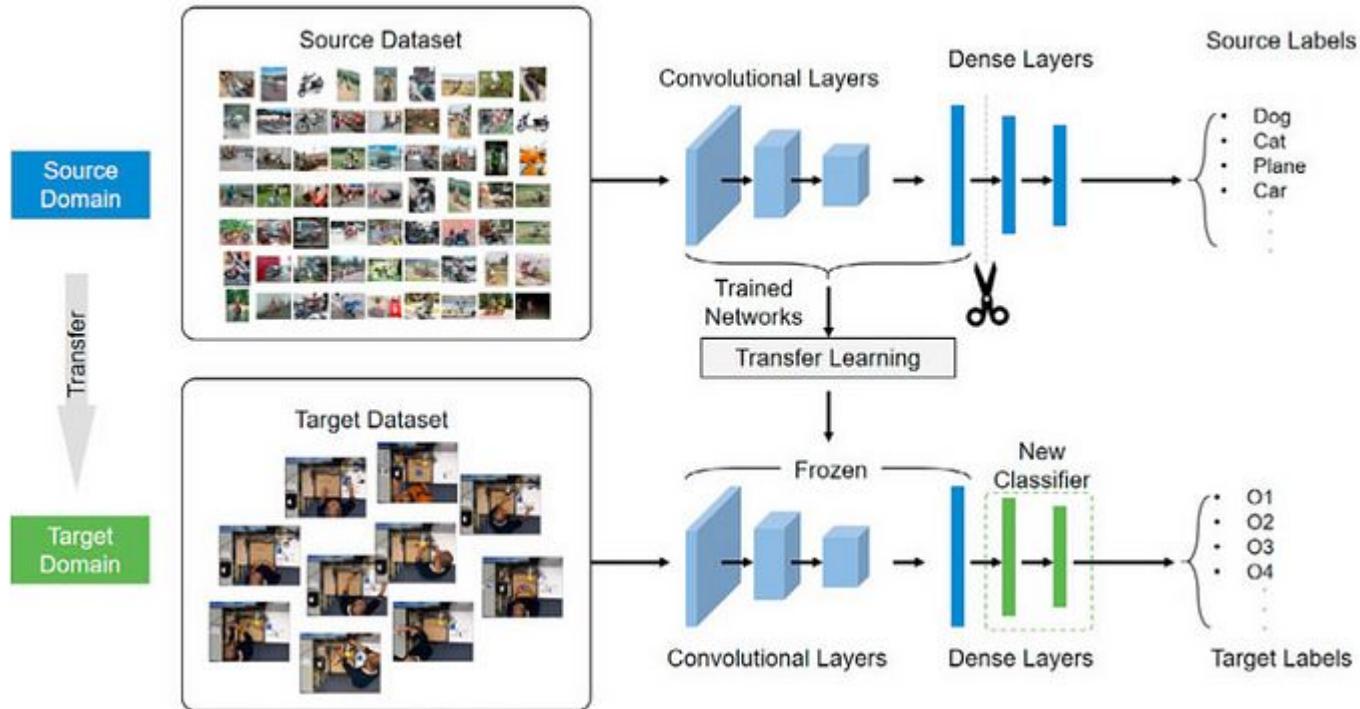
ایجاد Generalization به کمک روش‌های مختلف آموزش مدل

- Transfer learning
- Multi-task learning
- Self-supervised learning

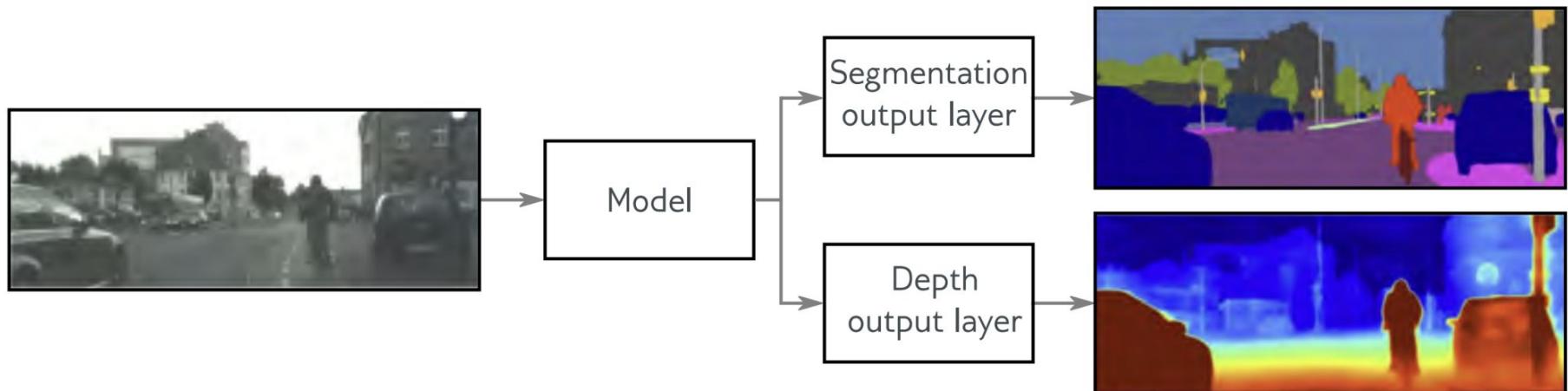
Transfer learning



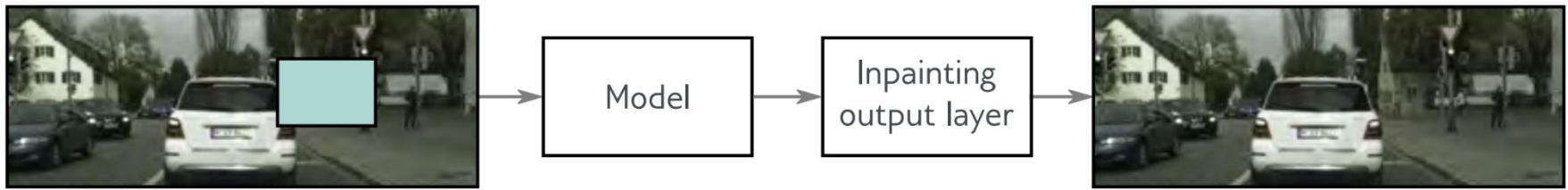
Transfer learning



Multi-task learning



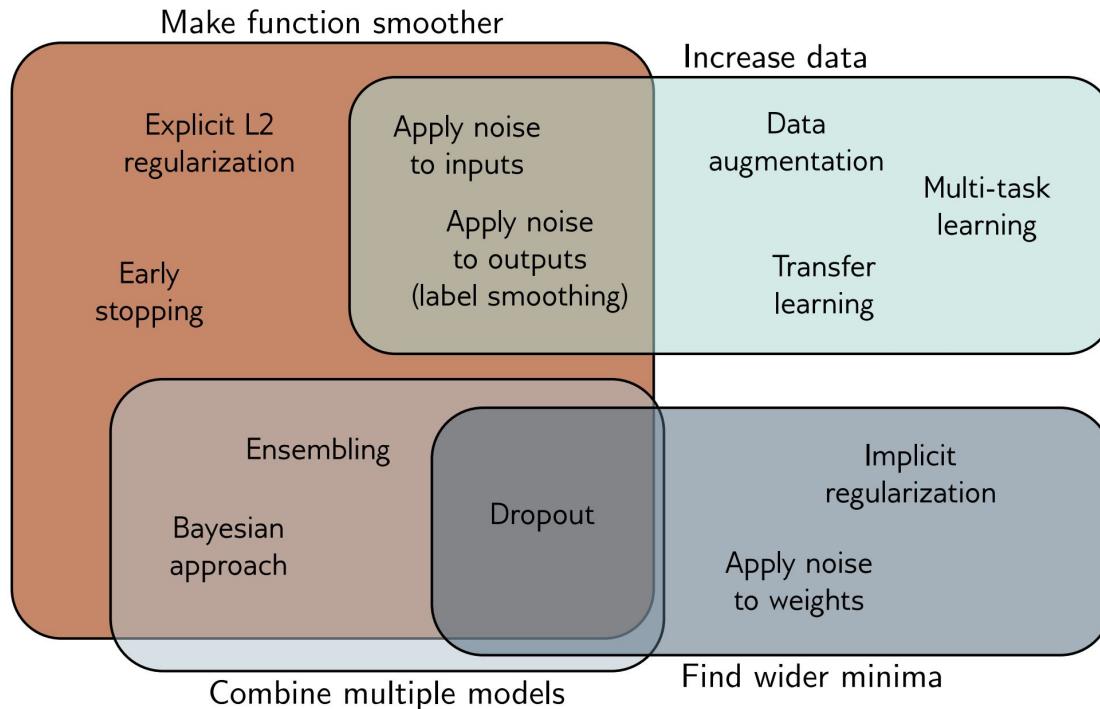
Self-supervised learning





جمع‌بندی روش‌های تعمیم‌دهی

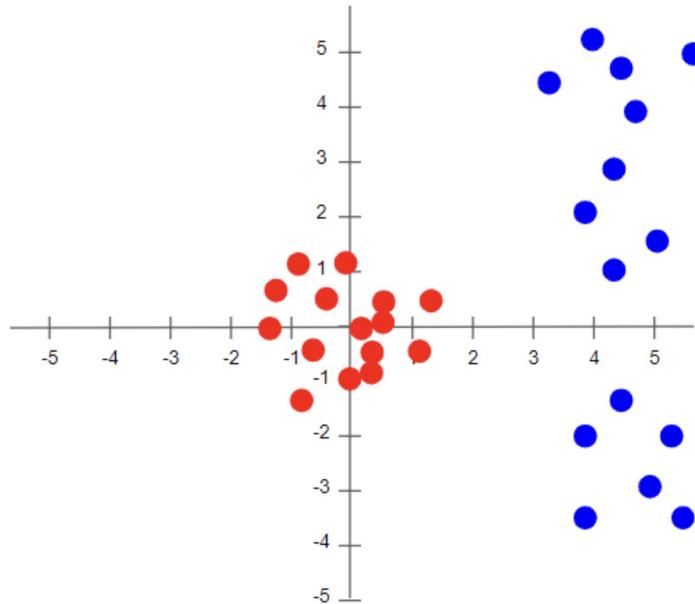
جمع‌بندی روش‌های تعمیم‌دهی



- Batch Normalization

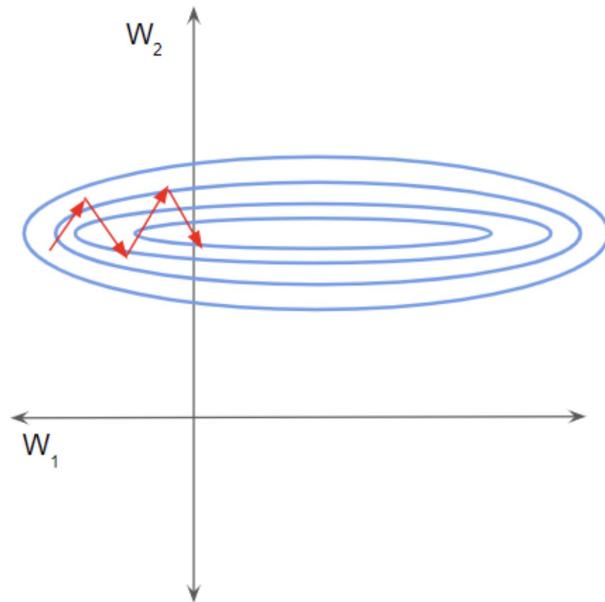
Batch Normalization

داده‌ها range های متفاوتی دارند.



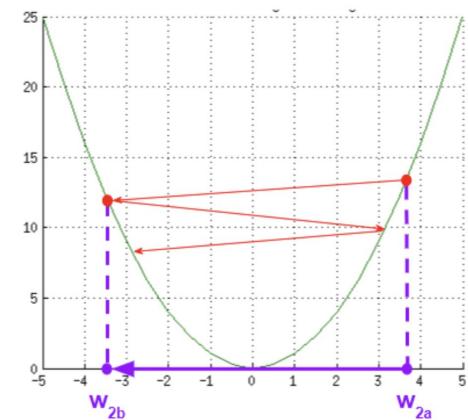
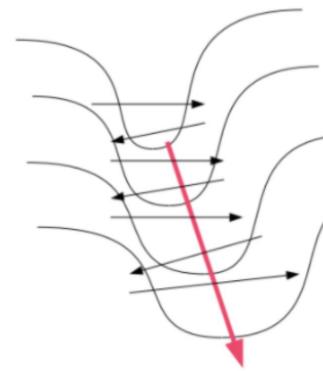
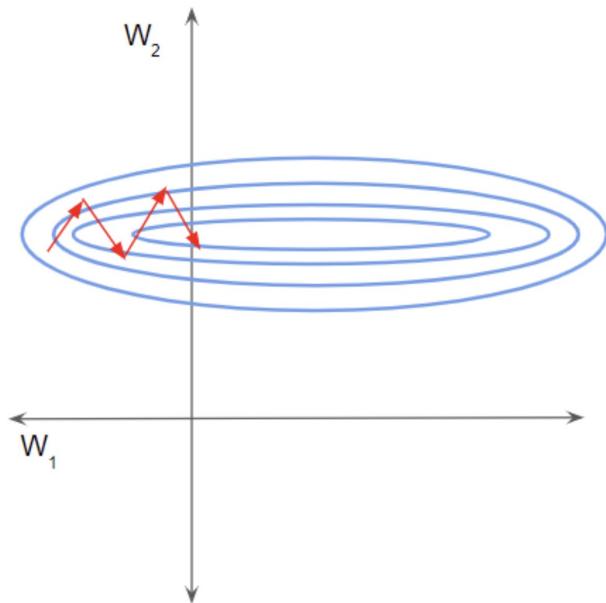
Batch Normalization

داده‌ها range های متفاوتی دارند.



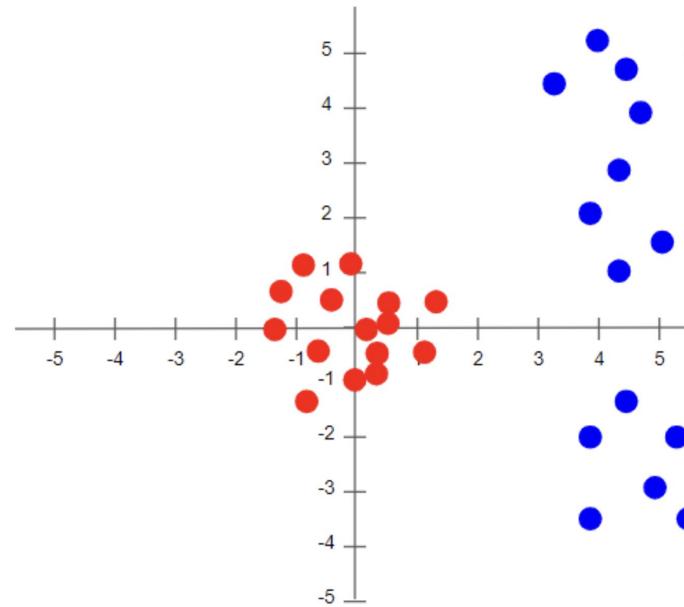
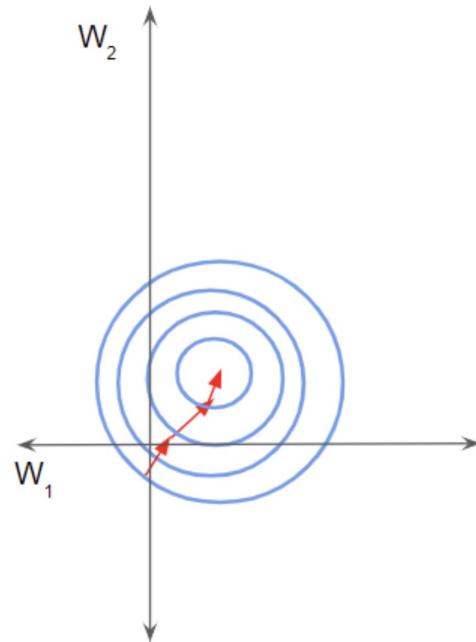
Batch Normalization

داده‌ها range‌های متفاوتی دارند.



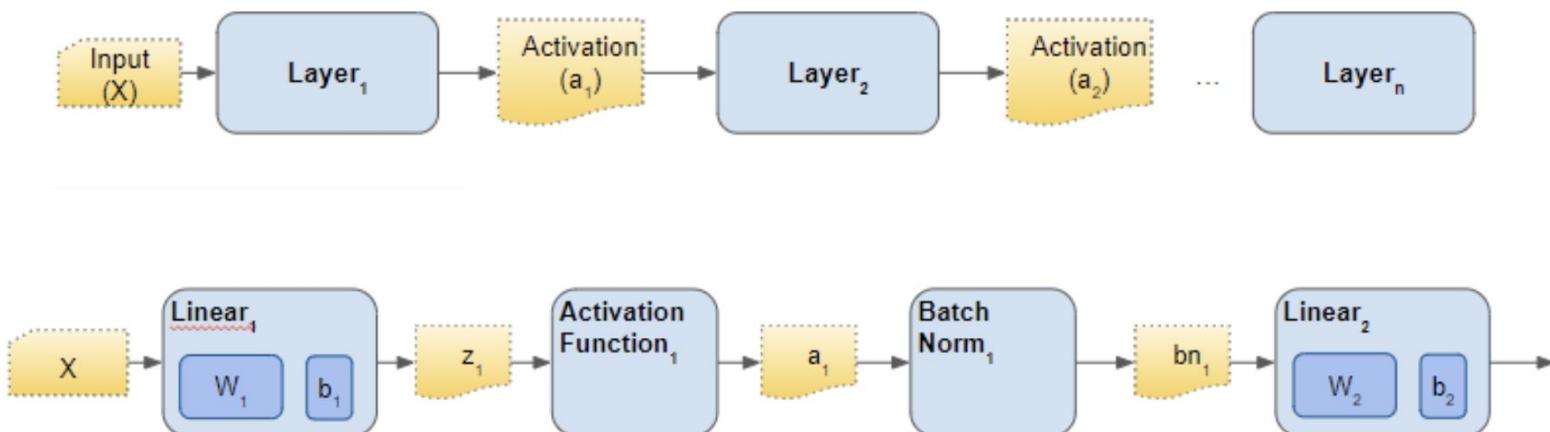
Batch Normalization

داده‌ها range‌های متفاوتی دارند.



Batch Normalization

محل اعمال در معماری مدل Batch Normalization



- Imbalanced Dataset

متريک‌های ارزیابی مدل

دقت (Accuracy) •

تعداد نمونه‌های درست پيش‌بینی شده

تعداد کل نمونه‌ها

كه يعني اگر از بين ۱۰۰ نمونه، ۹۸ تاى آن را درست پيش‌بینی کنيم، دقتمان ۹۸ درصد خواهد بود.

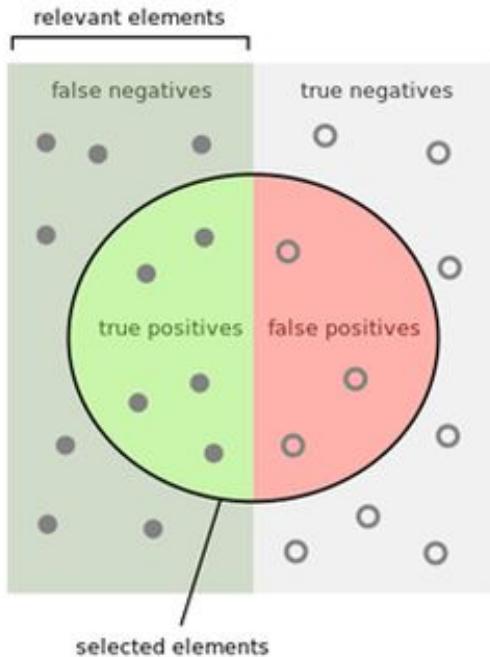
اين دقت خوب است؟

متريک‌های ارزیابی مدل

		Predicted		
		Positive	Negative	
Actual	Positive	True Positive	False Negative	Recall/Sensitivity $\frac{TP}{TP + FN}$
	Negative	False Positive	True Negative	Specificity $\frac{TN}{TN + FP}$
	Precision	$\frac{TP}{TP + FP}$	Negative Predictive Value $\frac{TN}{TN + FN}$	Accuracy $\frac{TP + TN}{TP + FN + TN + FP}$

- دقت (Accuracy)
- Recall
- Precision

متريک‌های ارزیابی مدل



How many selected items are relevant?

Precision =



How many relevant items are selected?

Recall =

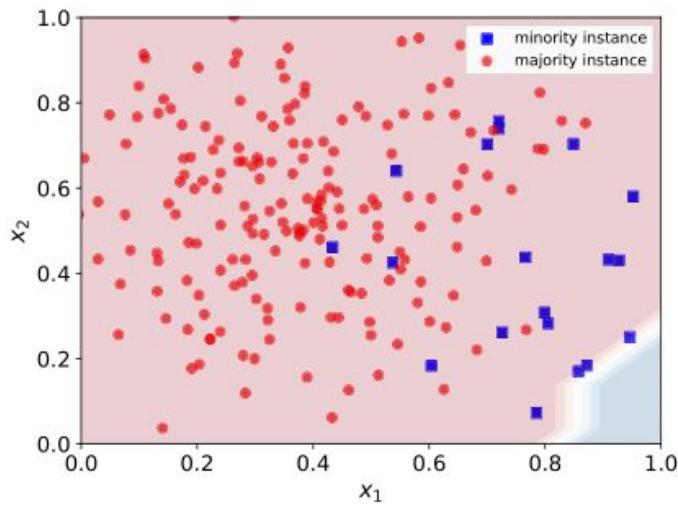
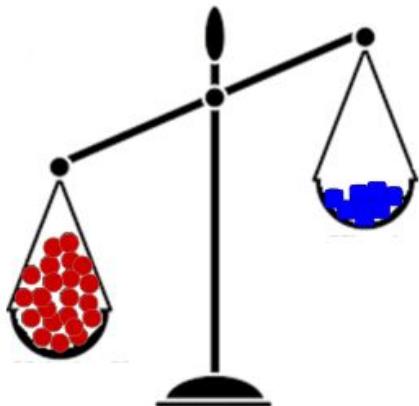


مجموعه داده‌های نامتوازن

- مجموعه داده‌های بیماران سرطانی و سالم

- مجموعه داده‌های داروهای تاثیرگذار بر روی یک بیماری خاص

- ...



مجموعه داده‌های نامتوابن

● مجموعه داده‌های بیماران سرطانی و سالم

● مجموعه داده‌های داروهای تاثیرگذار بر روی یک بیماری خاص

...

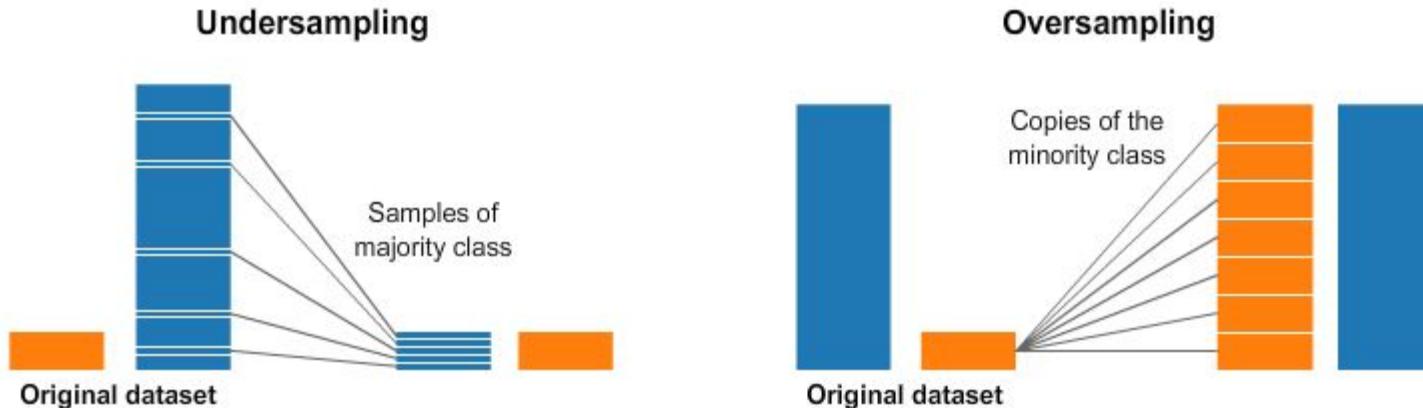
●

در تسکهای Regression و Classification به چه شکل می‌شود؟

راهکار مدیریت مجموعه داده‌های نامتوازن چیست؟

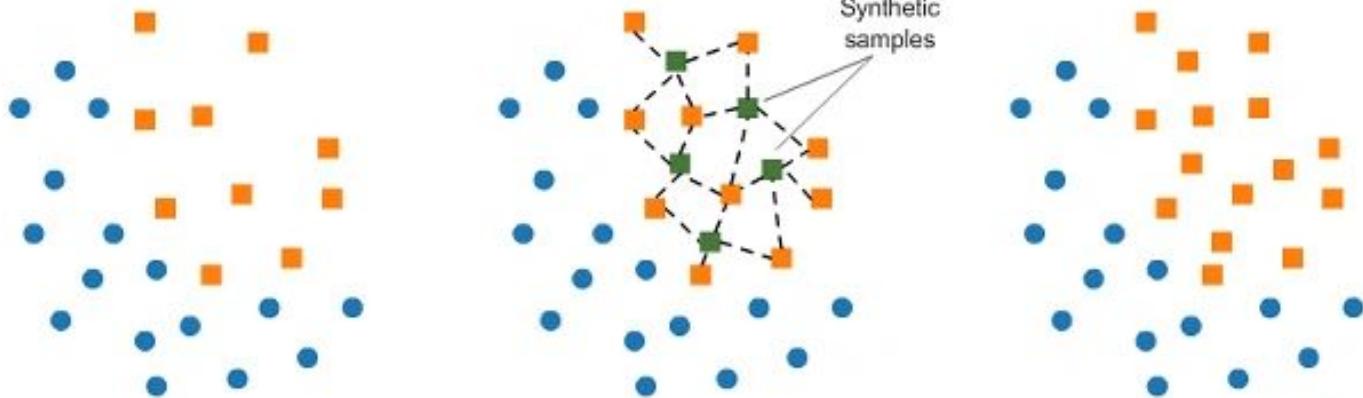
تغییر تعداد نمونه‌های هر کلاس

- انجام Oversampling برای داده‌های کلاس اقلیت
- انجام Undersampling برای داده‌های کلاس اکثریت



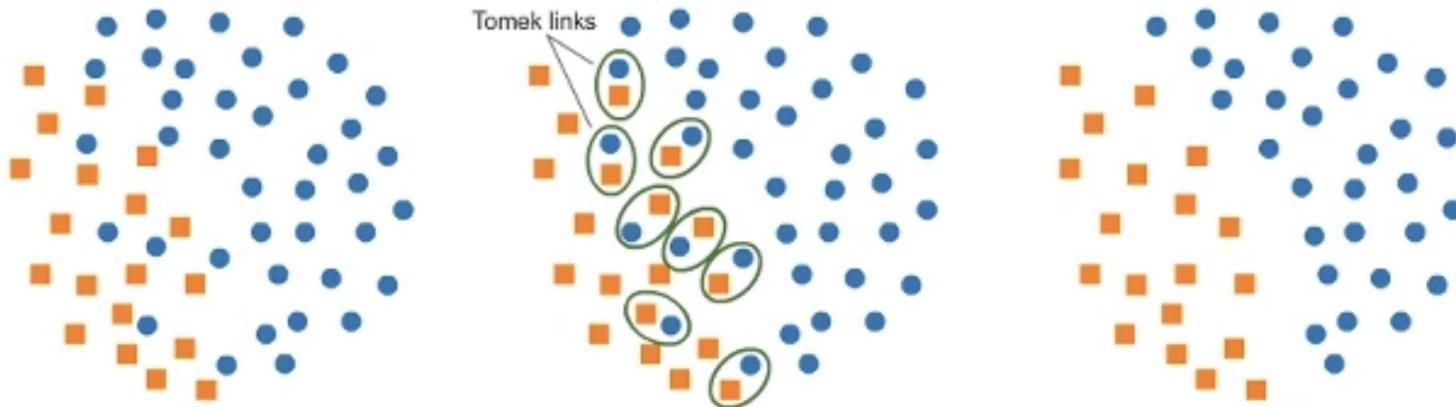
تغییر تعداد نمونه‌های هر کلاس

- تولید نمونه‌های مصنوعی



تغییر تعداد نمونه‌های هر کلاس

- حذف کردن هدفمند نمونه‌های کلاس اکثریت



تغییر Loss function

استفاده کردن از Weighted loss function به طوری که خطای پیش‌بینی کردن نمونه‌های مربوط

به کلاس اقلیت، هزینه‌ی بیشتری ایجاد کند.

$$\mathcal{L} = - (y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y}))$$

$$\mathcal{L} = - (w_1 \cdot y \cdot \log(\hat{y}) + w_0 \cdot (1 - y) \cdot \log(1 - \hat{y}))$$

تغییر Loss function

استفاده کردن از Weighted loss function به طوری که خطای پیش‌بینی کردن نمونه‌های مربوط



به کلاس اقلیت، هزینه‌ی بیشتری ایجاد کند.

$$\mathcal{L} = -(w_1 \cdot y \cdot \log(\hat{y}) + w_0 \cdot (1 - y) \cdot \log(1 - \hat{y})) \rightarrow \text{Binary Classification}$$

$$\mathcal{L} = - \sum_{c=1}^C w_c \cdot y_c \cdot \log(\hat{y}_c) \rightarrow \text{Multi-class Classification}$$

$$\mathcal{L}_{\text{weighted}} = \frac{1}{\sum_{i=1}^N w_i} \sum_{i=1}^N w_i \cdot (y_i - \hat{y}_i)^2 \rightarrow \text{Regression}$$

تغییر Loss function

استفاده کردن از Weighted loss function به طوری که خطای پیش‌بینی کردن نمونه‌های مربوط

به کلاس اقلیت، هزینه‌ی بیشتری ایجاد کند.

$$\mathcal{L} = - \sum_{c=1}^C w_c \cdot y_c \cdot \log(\hat{y}_c) \quad \rightarrow \quad \text{Multi-class Classification}$$

$$w_c = \frac{N}{n_c}$$

Class	Number of examples	Weight
A	1,000	1.01
B	10	101

تغییر Loss function

استفاده کردن از Weighted loss function به طوری که خطای پیش‌بینی کردن نمونه‌های مربوط

به کلاس اقلیت، هزینه‌ی بیشتری ایجاد کند.

$$\mathcal{L} = -(w_1 \cdot y \cdot \log(\hat{y}) + w_0 \cdot (1 - y) \cdot \log(1 - \hat{y}))$$

$$\mathcal{L} = - \sum_{c=1}^C w_c \cdot y_c \cdot \log(\hat{y}_c) \quad w_c = \frac{N}{n_c}$$

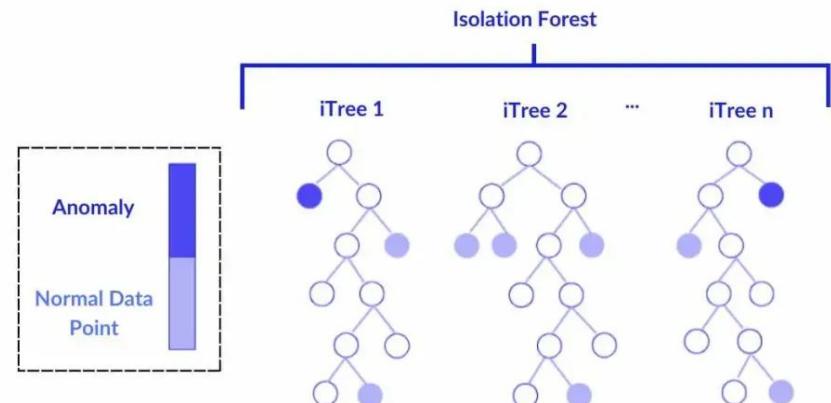
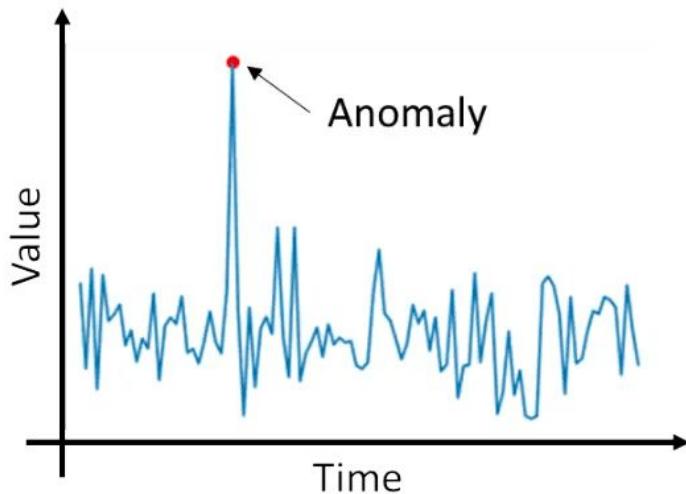
سایر روش‌های مدیریت داده‌ی نامتوازن

استفاده کردن از مدل‌های مناسب این نوع داده‌ها. به طور مثال مدل‌های ensemble

-

Anomaly detection

-





جمع بندی

جمع‌بندی

● با روش‌های مختلف تعمیم داده آشنا شدیم:

- Regularization
- Model ensembles
- Dropout

● با انواع روش‌های مدیریت داده‌های نامتوارن آشنا شدیم:

- Resampling
- Weighted loss function
- ...

پایان