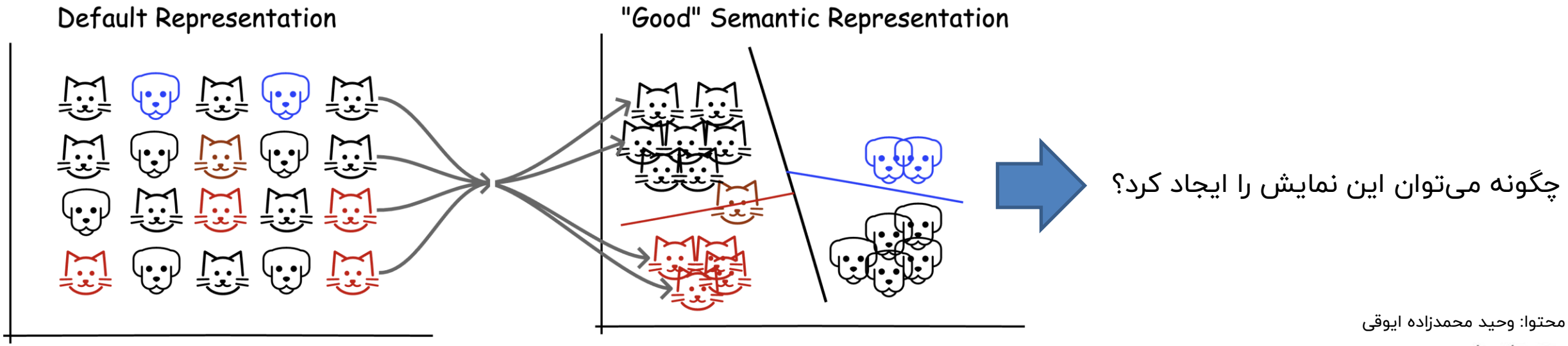
Representation Learning

dayche.com . گروه دایکه

دای‌کِه DAYCHE

Success of machine learning algorithms

- عملکرد الگوریتم‌های یادگیری ماشین وابستگی مستقیم به نمایش داده‌ها دارد.
  - این حقیقت مبتنی بر عملکرد انسان نیز است.
- بازنمایی در حوزه یادگیری ماشین
  - بهترین بازنمایی از داده‌ها به منظور انجام یک وظیفه در ساده‌ترین حالت ممکن

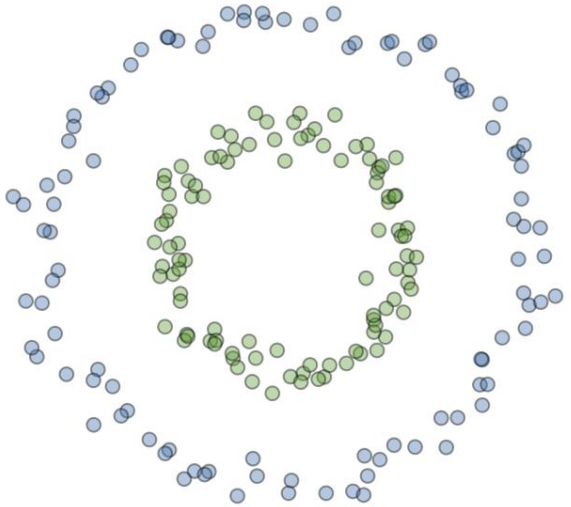


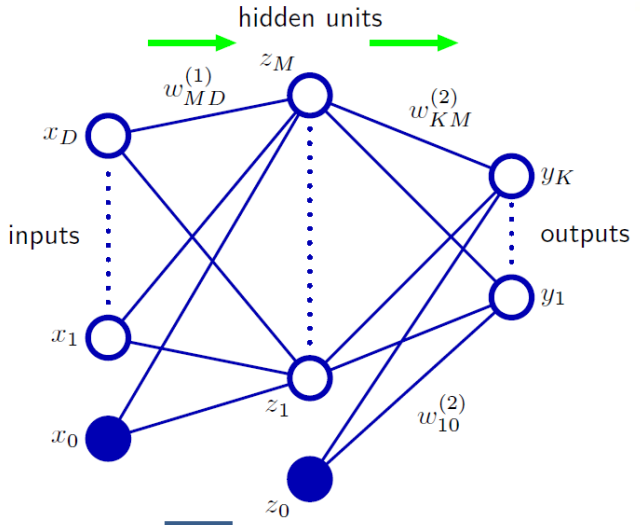
چگونه می‌توان این نمایش را ایجاد کرد؟

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

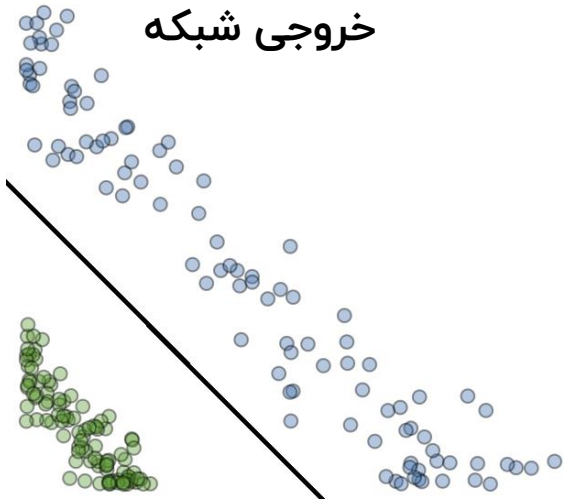
Feed forward neural networks

ورودی شبکه

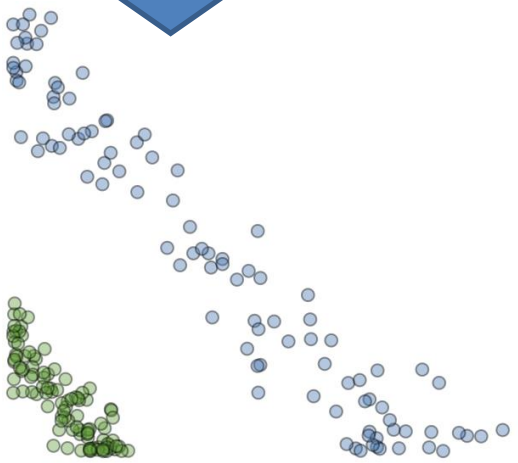




خروجی شبکه



آیا لایه آخر هر مدلی، به عنوان مثال ماشین بردار پشتیبان، می‌تواند باشد؟



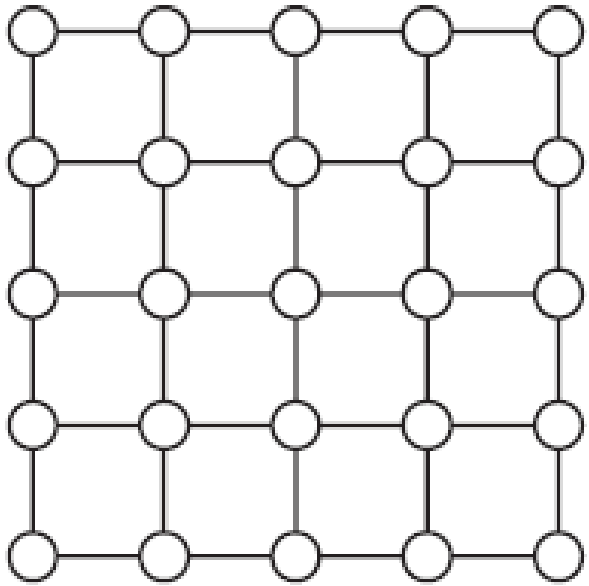
خروجی لایه مخفی

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

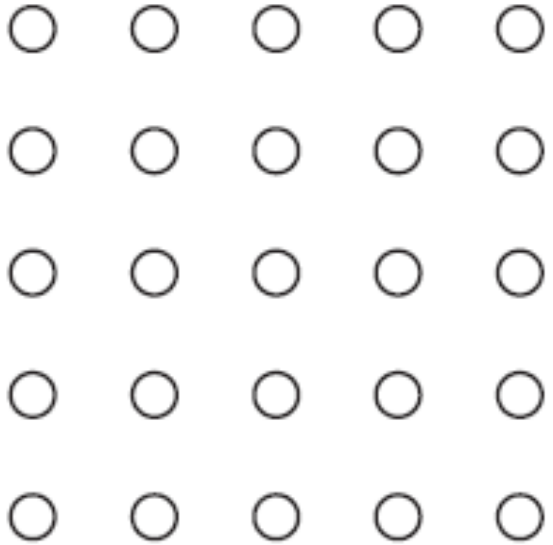
Constrains on code layer

- نمایش داده‌ها به شکل نشان داده شده به صورت صریح از شبکه خواسته نشده است.
  - شبکه سعی می‌کند معیار آموزشی خواسته شده را برآورده کند.
- برخی از مدل‌ها به صورت صریح آموزش می‌بینند تا ویژگی‌های مشخصی را برآورده کنند.
  - تسک تخمین تابع چگالی احتمال

فضای اصلی



فضای مطلوب

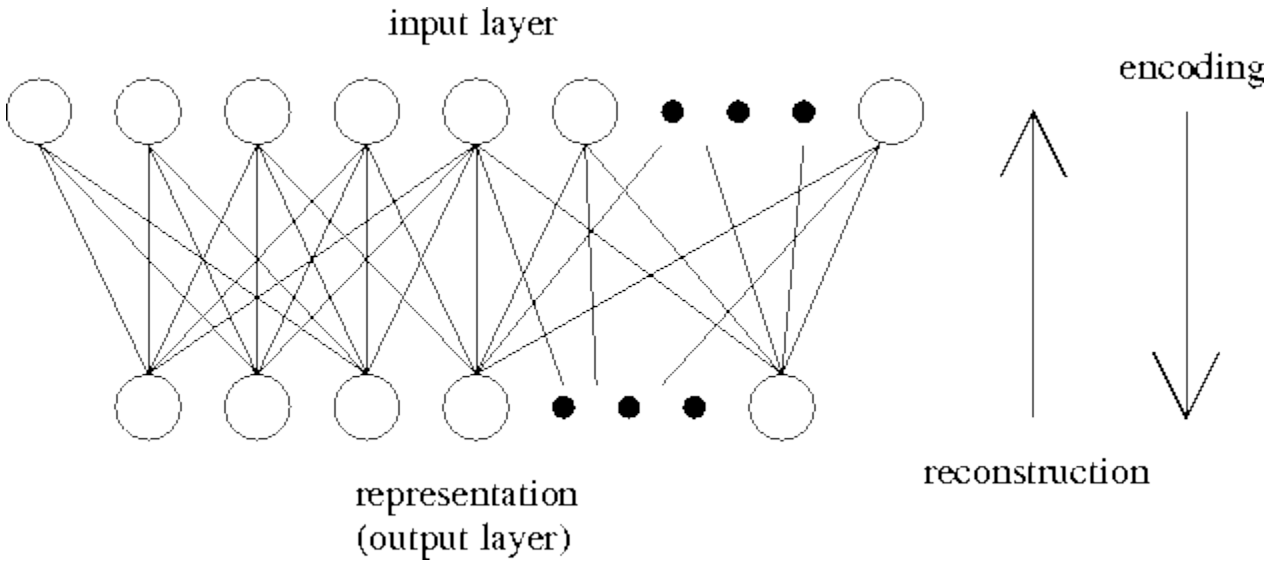


تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Representation learning

- برای یادگیری بازنمایی دو قاعده کلی وجود دارد:
  - حفظ بیشینه اطلاعات – خطای بازیابی
  - استخراج ویژگی‌های مناسب

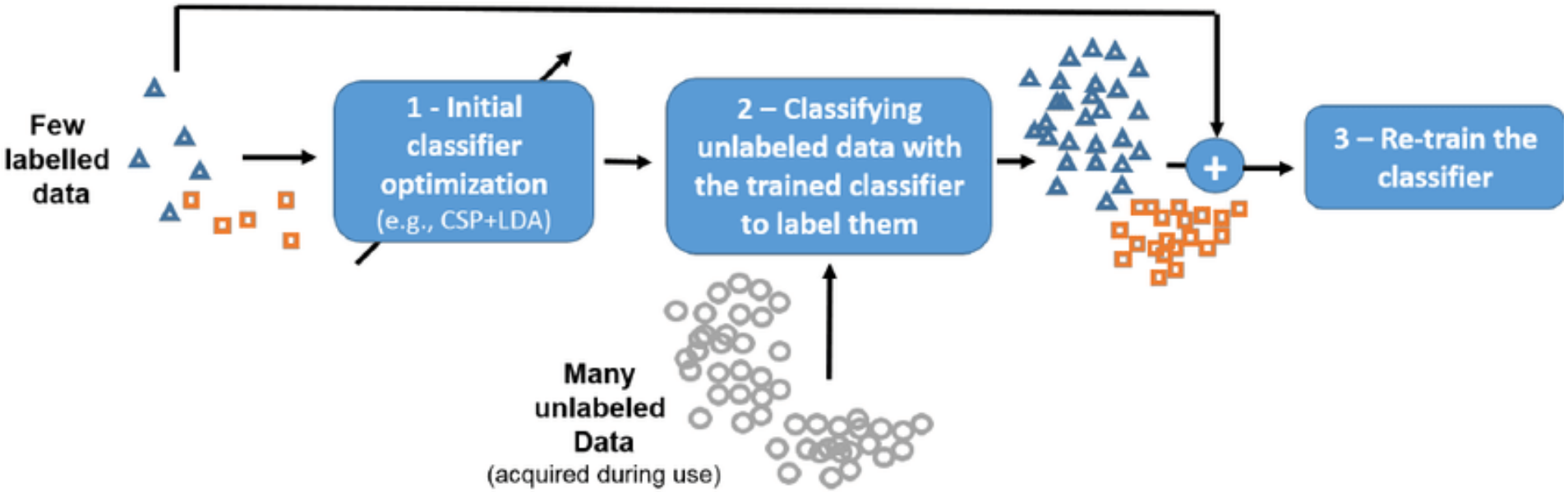
- آموزش بدون نظارت
- یادگیری بازنمایی زمینه آموزش نیمه نظارتی را نیز فراهم می‌کند
- یادگیری انتقال



تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Semi-supervised learning

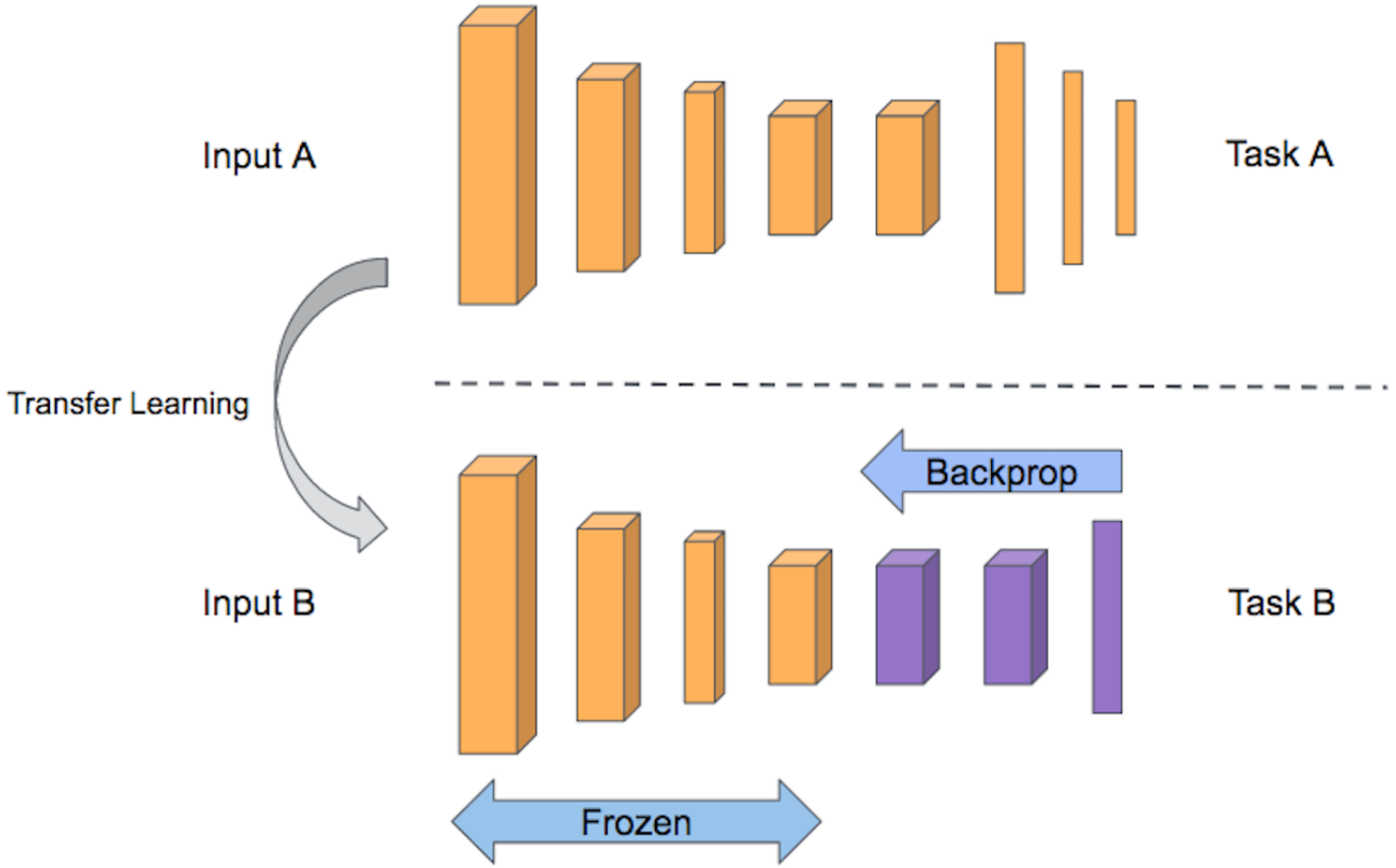
- یادگیری نیمه نظارتی



تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Transfer learning

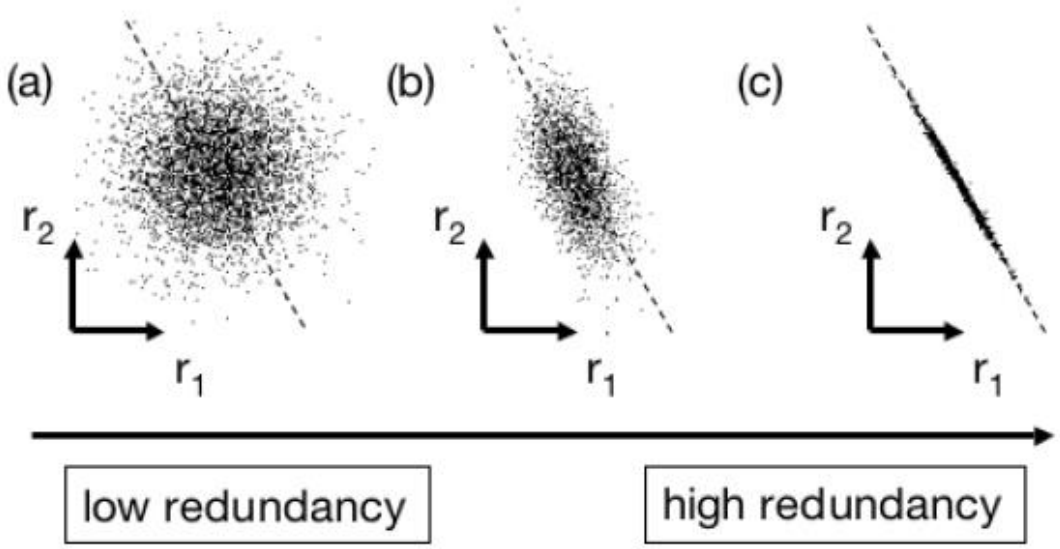
- یادگیری انتقال



تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

PCA

- آنالیز مولفه‌های اصلی
  - حذف اطلاعات اضافی به منظور کاهش محاسبات



$$Y = PX$$

چطور می‌توان ماتریس کوواریانس Y را قطری کرد؟

$$YY^T = PXX^T P^T \rightarrow \Sigma_y = P\Sigma_x P^T$$

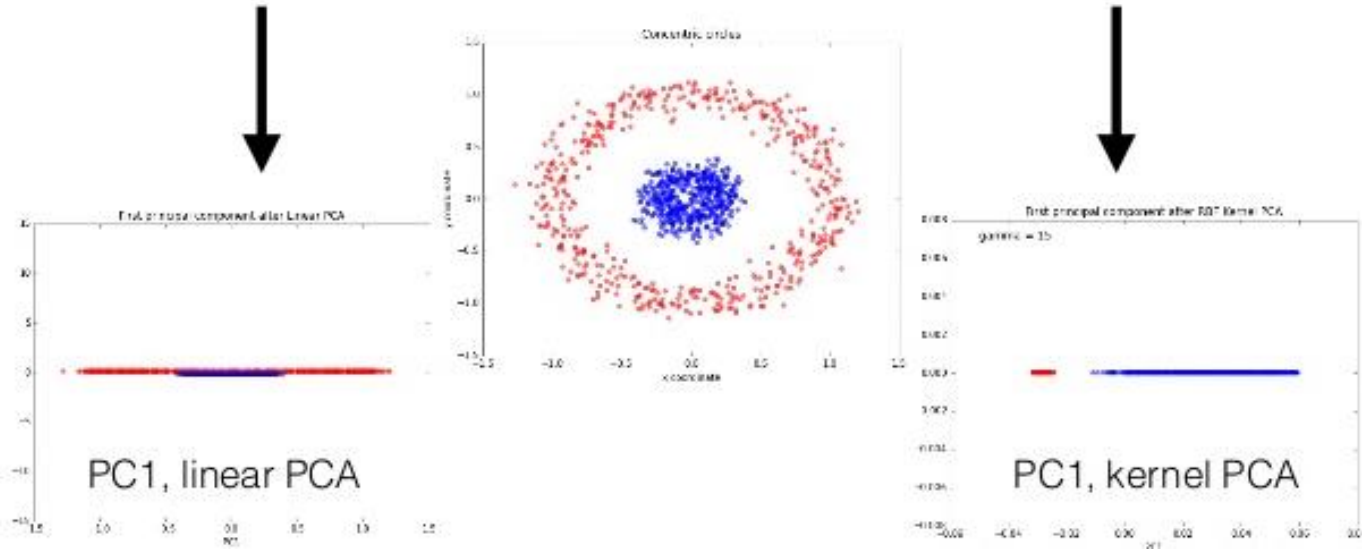
تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Kernel PCA

- آنالیز مولفه‌های اصلی مبتنی بر کرنل
  - ایراد روش خطی

$$\text{Cov} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i^T \boldsymbol{x}_i$$

$$\text{Cov} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\phi}(\boldsymbol{x}_i)^T \boldsymbol{\phi}(\boldsymbol{x}_i)$$



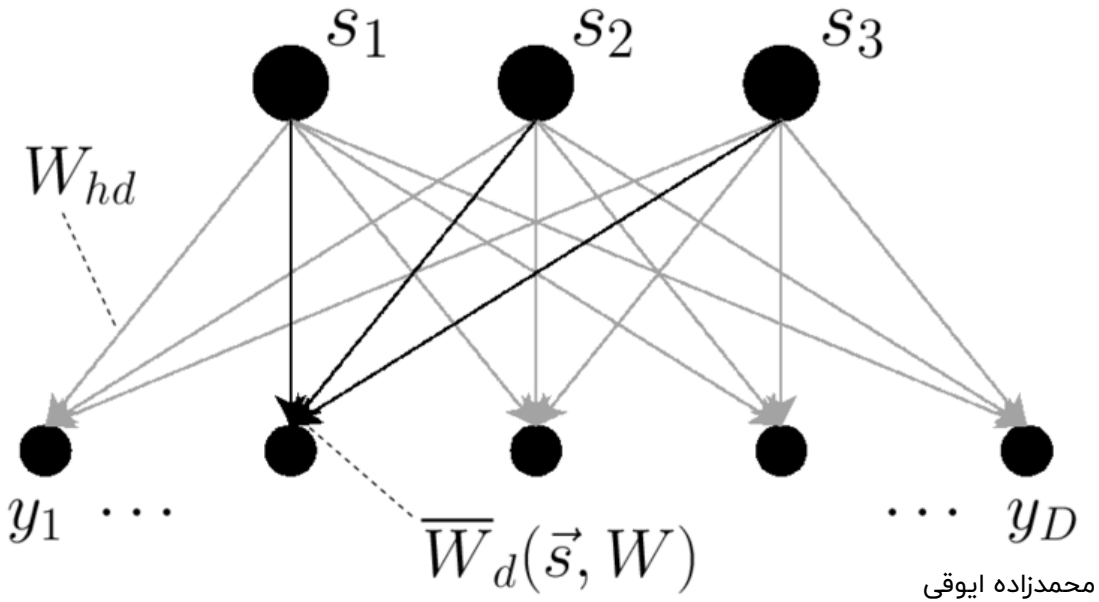
تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Representation learning approaches

- دو روند کلی برای یادگیری بازنمایی وجود دارد:
  - مدل‌های گرافی احتمالاتی
  - شبکه‌های عصبی عمیق

پیدا کردن یک تابع که قادر به مدل کردن تغییرات موجود در داده‌ها است.
- به دلیل وجود وابستگی در داده‌ها، به صورت مستقیم نمی‌توان با آنها کار کرد.
- از دیدگاه ریاضی نیاز به تغییر در نمایش داده‌ها داریم

مدل‌های یادگیر بازنمایی، دارای یک لایه متغیر مخفی است



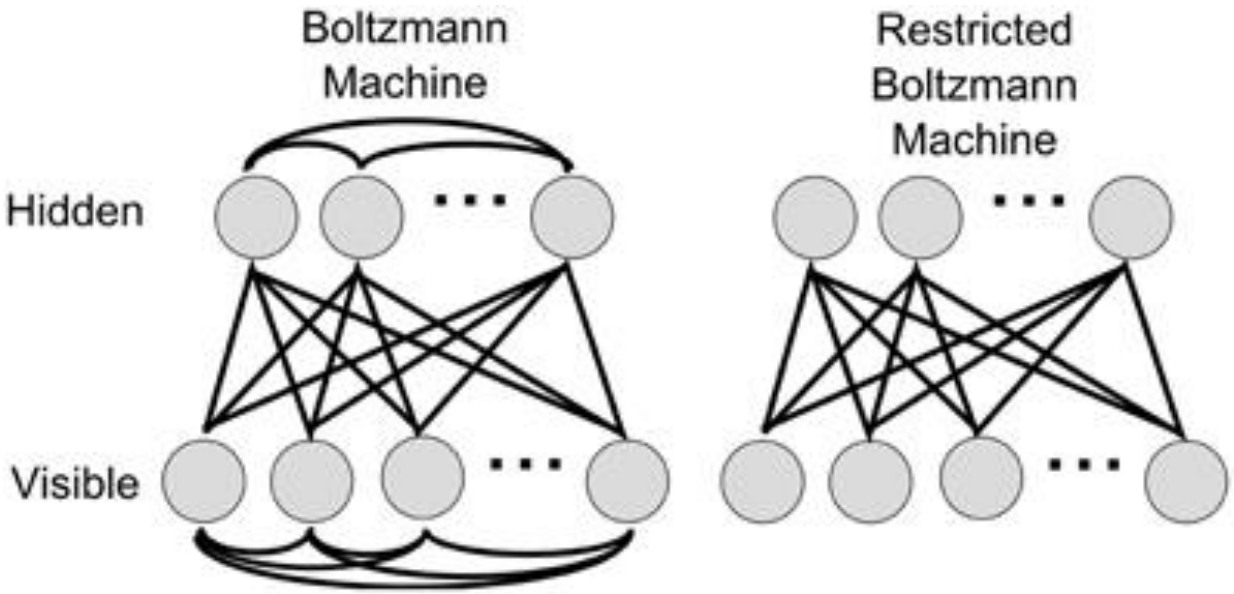
تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Probabilistic graphical models

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Probabilistic graphical models

- مدل‌های مولد (Generative models)
- ماشین بولتزمن

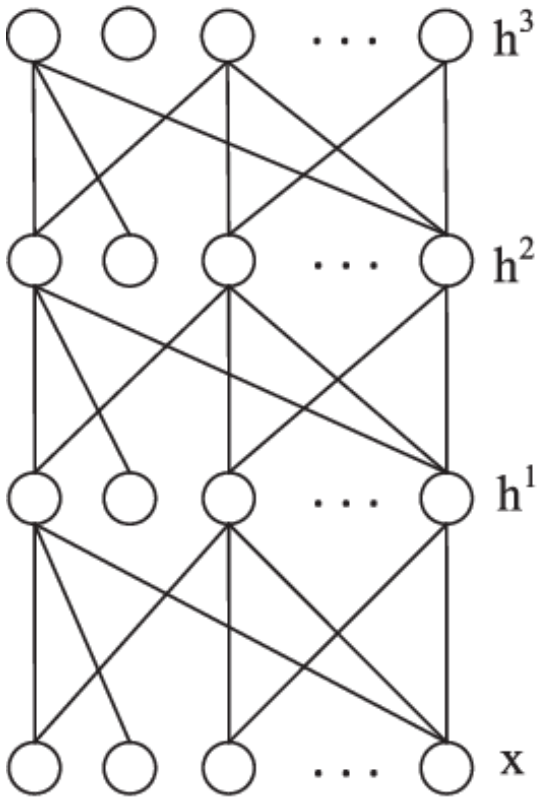


مدل مبتنی بر انرژی – توسعه یافته شده توسط Hinton
- رویکرد غیراحتمالاتی
- آموزش بر اساس تقریب تابع بیشینه شباهت

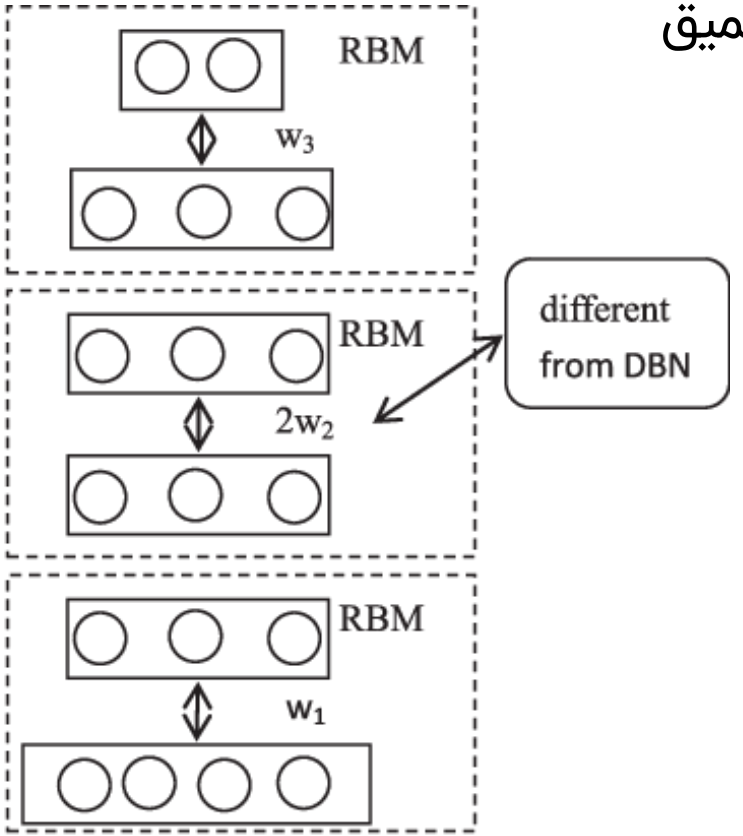
تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
گروه دایکه | dayche.com

Probabilistic graphical models

- ماشین بولتزمن عمیق



Deep Boltzmann Machine

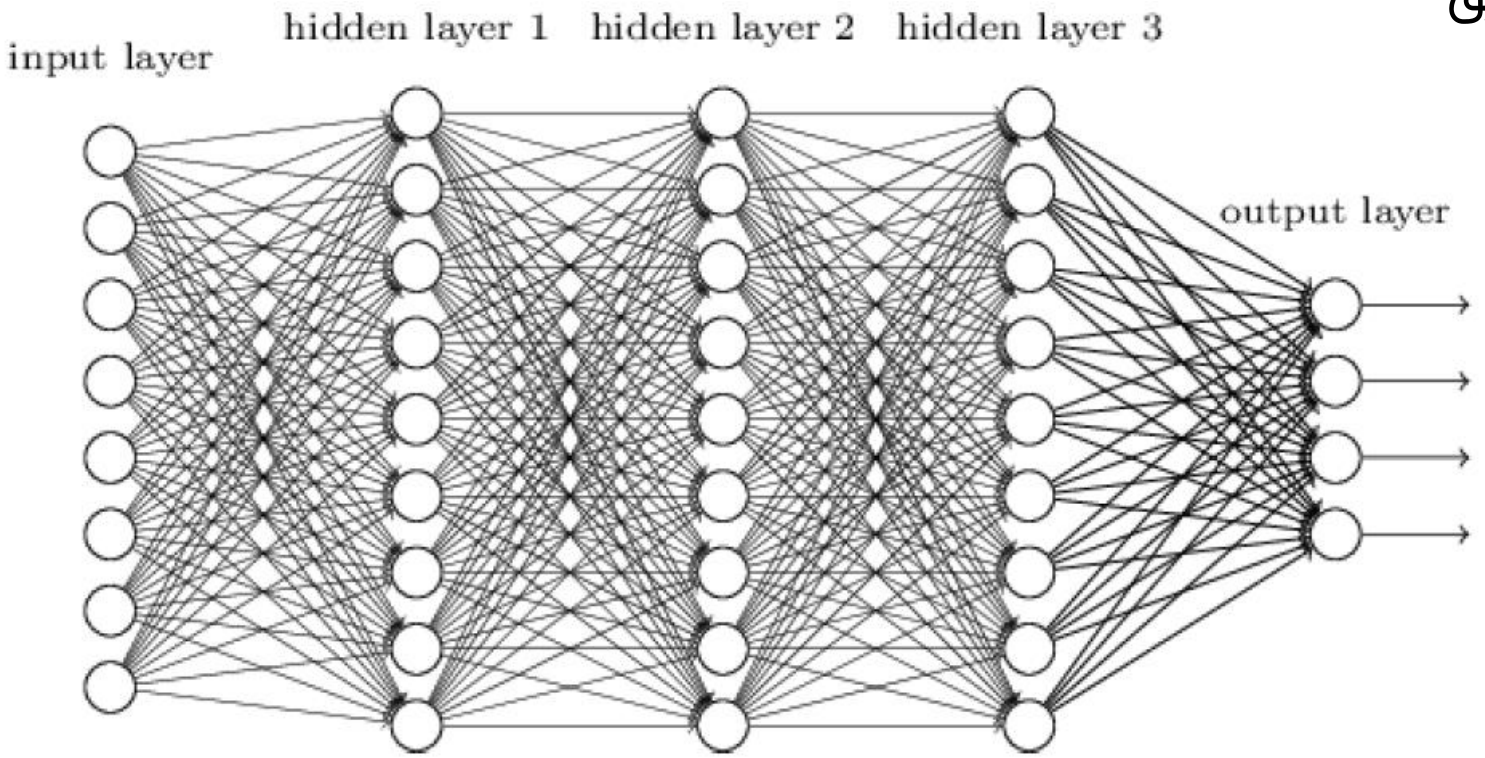


Pre-training of DBM

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Probabilistic graphical models

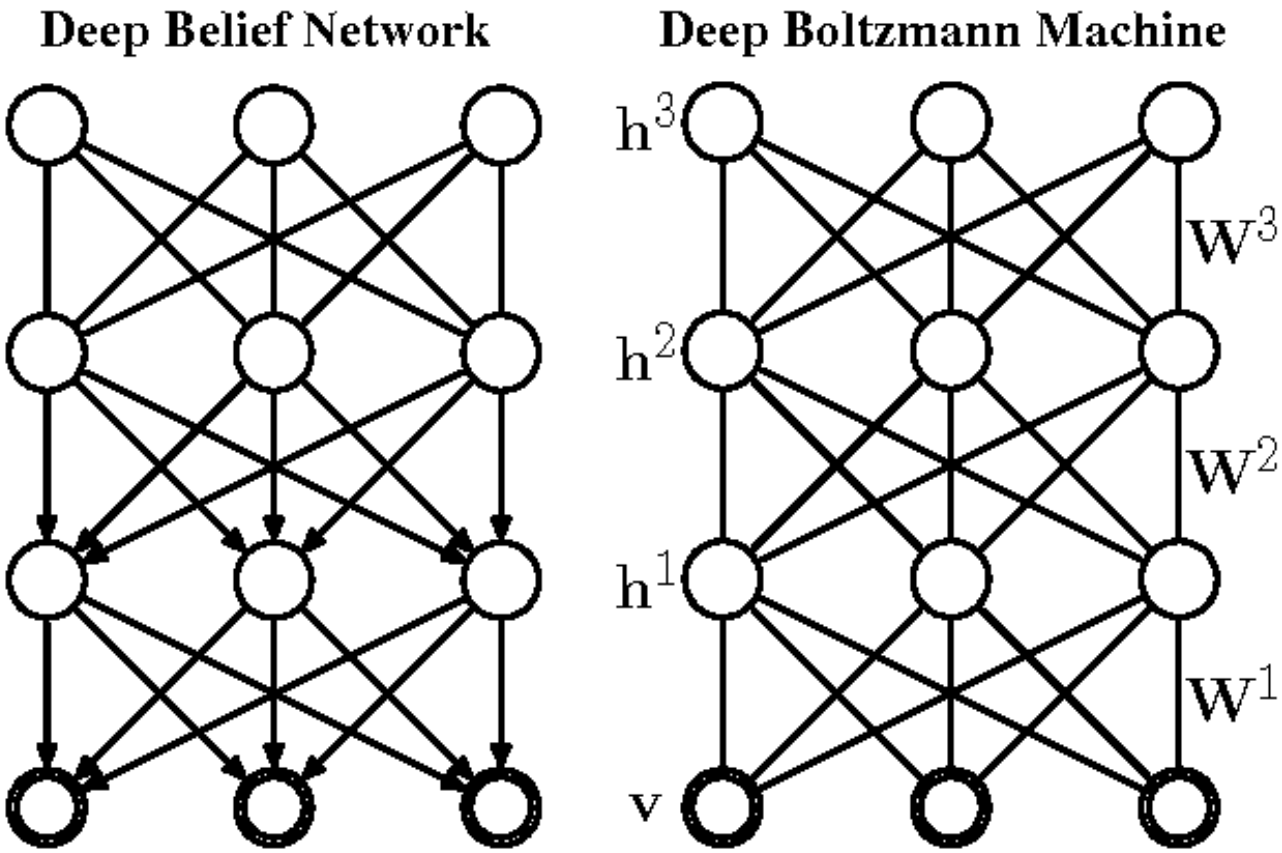
- شبکه باور عمیق



تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
گروه دایکه | dayche.com

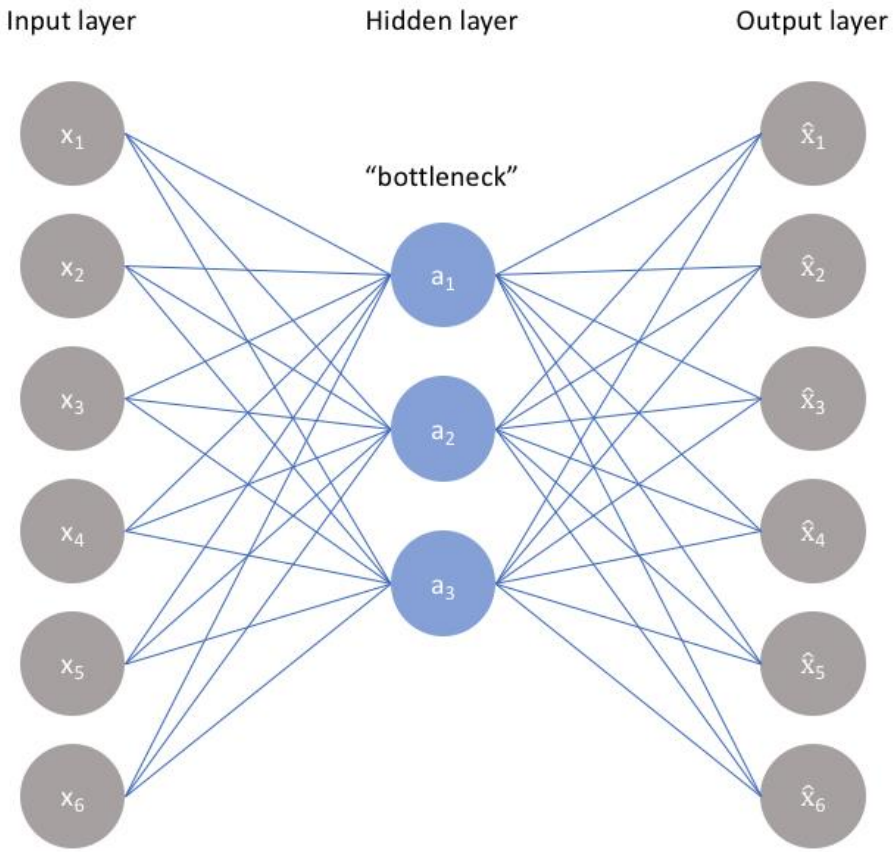
Probabilistic graphical models

- تفاوت شبکه باور عمیق و ماشین بولتزمن عمیق



تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Auto-encoders



- کدکننده‌های خودکار Auto-encoders
- یادگیری بدون نظارت

آموزش بر اساس خطای بازیابی

$$L(\theta) = |x - \hat{x}|_2^2$$

$$net_{en} = W_{en}X + b_{en}, \qquad o_{en} = f_{en}(net_{en})$$

$$net_{de} = W_{de}o_{en} + b_{de}, \qquad o_{de} = \underbrace{f_{de}}_{\text{معمولا خطی}}(net_{de})$$

اگر لایه کدکننده نیز خطی باشد، نتیجه به چه صورت خواهد بود؟

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Auto-encoders training

$$J = \frac{1}{2}E^T E, \qquad E = x - \hat{x}$$

روابط مسیر پیشرو:
$$\begin{cases} net_{en} = W_{en}X + b_{en}, \quad o_{en} = f_{en}(net_{en}) \\ net_{de} = W_{de}o_{en} + b_{de}, \hat{x} = net_{de} \end{cases}$$

روابط مسیر برگشت:

$$\frac{\partial J}{\partial W^{de}} = \frac{\partial J}{\partial E} \times \frac{\partial E}{\partial \hat{x}} \times \frac{\partial \hat{x}}{\partial net_{de}} \times \frac{\partial net_{de}}{\partial W^{de}} = E \times -1 \times 1 \times O_{en} = -EO_{en}$$

$$\frac{\partial J}{\partial b^{de}} = \frac{\partial J}{\partial E} \times \frac{\partial E}{\partial \hat{x}} \times \frac{\partial \hat{x}}{\partial net_{de}} \times \frac{\partial net_{de}}{\partial b_{de}} = E \times -1 \times 1 = -E$$

$$\frac{\partial J}{\partial W^{en}} = \frac{\partial J}{\partial E} \times \frac{\partial E}{\partial \hat{x}} \times \frac{\partial \hat{x}}{\partial net_{de}} \times \frac{\partial net_{de}}{\partial O^{en}} \times \frac{\partial O^{en}}{\partial net_{en}} \times \frac{\partial net_{en}}{\partial W^{en}} = E \times -1 \times 1 \times W^{de} \times f^{en'}(net_{en}) \times X = -EW^{de}f^{en'}(net_{en})X$$

$$\frac{\partial J}{\partial b^{en}} = \frac{\partial J}{\partial E} \times \frac{\partial E}{\partial \hat{x}} \times \frac{\partial \hat{x}}{\partial net_{de}} \times \frac{\partial net_{de}}{\partial O^{en}} \times \frac{\partial O^{en}}{\partial net_{en}} \times \frac{\partial net_{en}}{\partial b^{en}} = E \times -1 \times 1 \times W^{de} \times f^{en'}(net_{en}) \times 1 = -EW^{de}f^{en'}(net_{en})$$

تولید محتوا: وحید محمدزاده ایوقی

daychegroup
daychegroup
گروه دایکه | dayche.com

Parameters sharing

- عملکرد کدکننده و کدگشا به شدت وابسته به آموزش پارامترها است.
  - اشتراک پارامترها

- کاهش پارامترهای مدل – کاهش احتمال بیش برازش شدن
- افزایش سرعت آموزش

روابط مسیر برگشت:

$$\frac{\partial J}{\partial W^{de}} = \frac{\partial J}{\partial E} \times \frac{\partial E}{\partial \hat{x}} \times \frac{\partial \hat{x}}{\partial net_{de}} \times \frac{\partial net_{de}}{\partial W^{de}} = E \times -1 \times 1 \times O_{en} = -EO_{en}$$

$$\frac{\partial J}{\partial b^{de}} = \frac{\partial J}{\partial E} \times \frac{\partial E}{\partial \hat{x}} \times \frac{\partial \hat{x}}{\partial net_{de}} \times \frac{\partial net_{de}}{\partial b_{de}} = E \times -1 \times 1 = -E$$

$$W_{en} = W_{de}^{T}$$

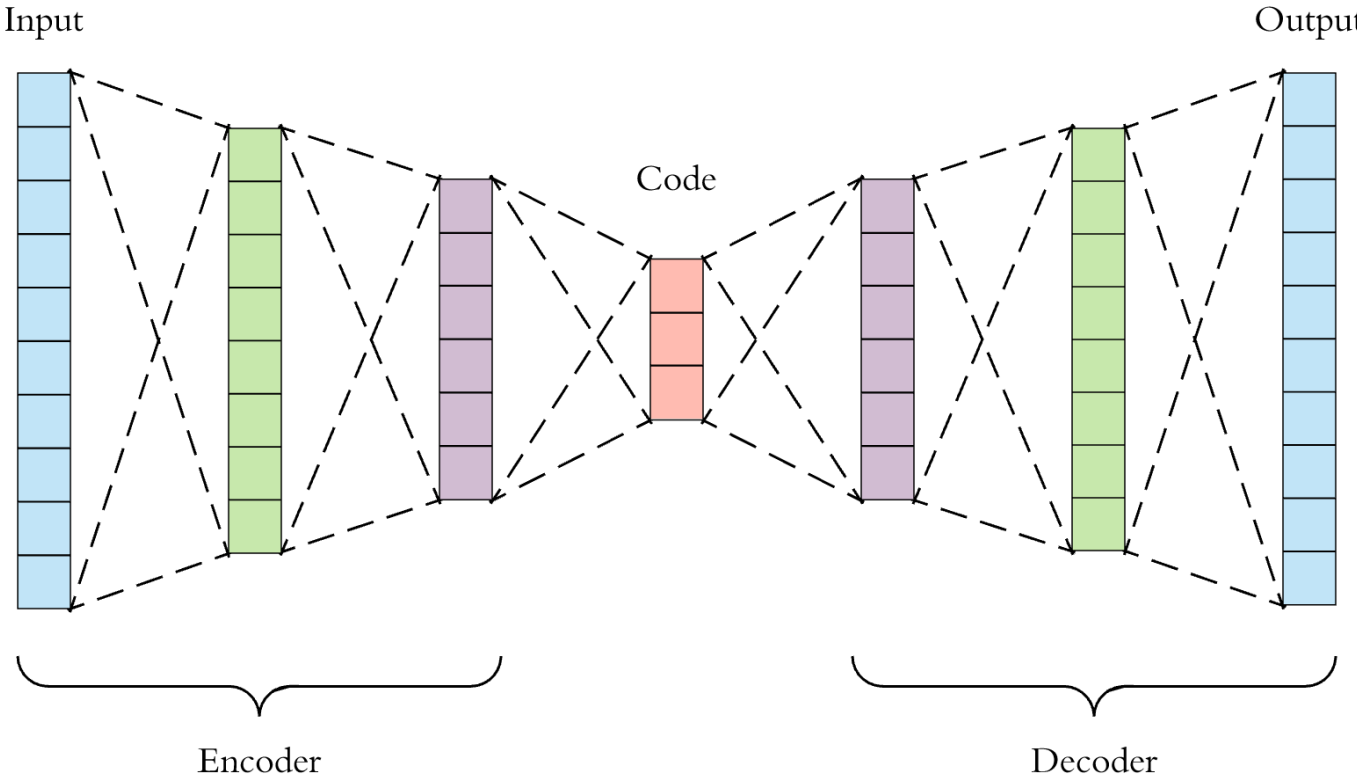
$$\frac{\partial J}{\partial b^{en}} = \frac{\partial J}{\partial E} \times \frac{\partial E}{\partial \hat{x}} \times \frac{\partial \hat{x}}{\partial net_{de}} \times \frac{\partial net_{de}}{\partial O^{en}} \times \frac{\partial O^{en}}{\partial net_{en}} \times \frac{\partial net_{en}}{\partial b^{en}} = E \times -1 \times 1 \times W^{de} \times f^{en\prime}(net_{en}) \times 1 = -EW^{de} f^{en\prime}(net_{en})$$

تولید محتوا: وحید محمدزاده ایوقی

daychegroup
daychegroup
گروه دایکه | dayche.com

Deep auto-encoders

- کدکننده‌های خودکار عمیق Deep Auto-encoders
  - شبکه‌های عمیق دارای قدرت بازنمایی بالاتری نسبت به شبکه‌های کم عمق هستند.

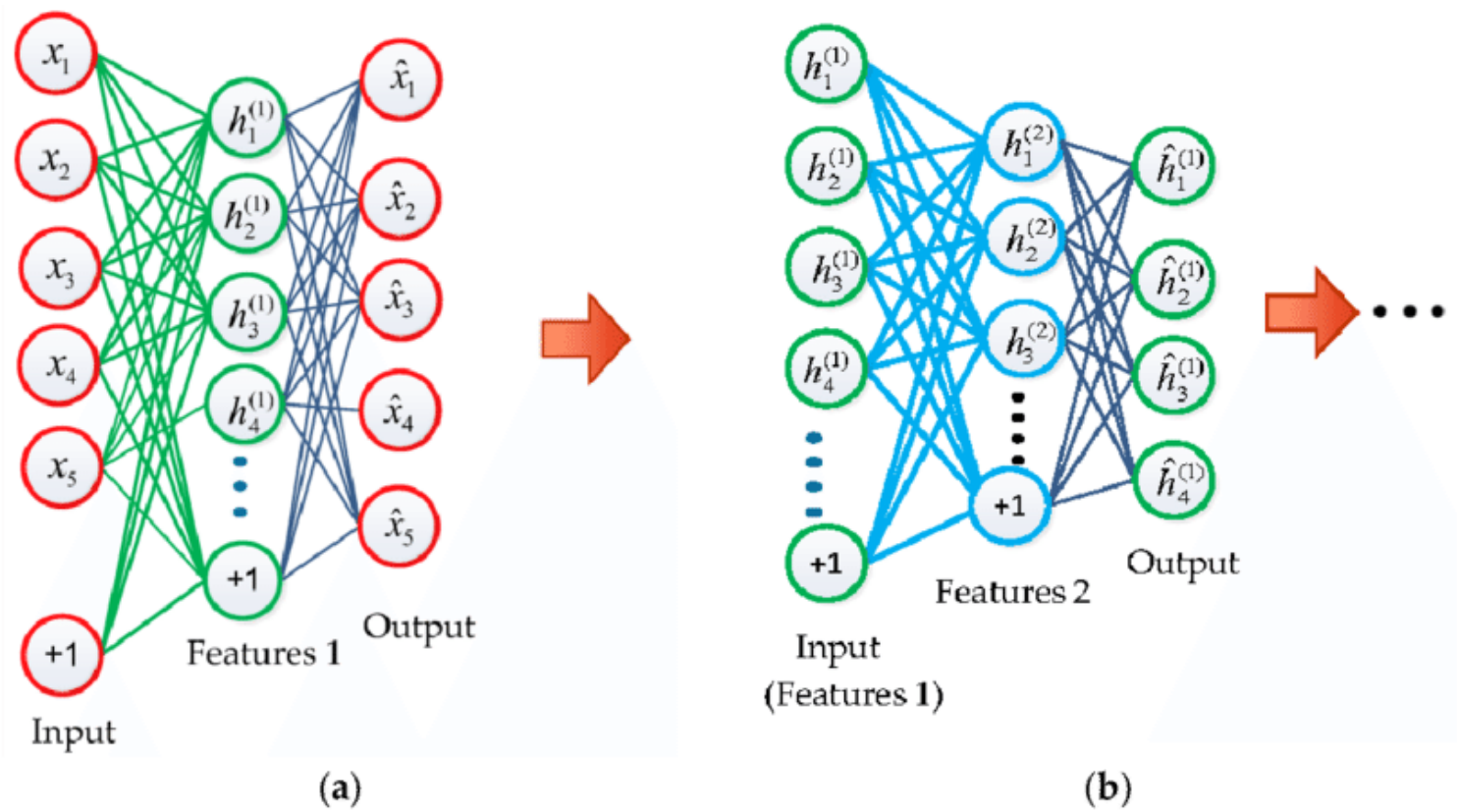


مشکل آموزش – محو شدگی گرادیان
راه حل
- آموزش لایه به لایه بدون نظارت
- آموزش لایه به لایه با نظارت

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Unsupervised layer-wised pre-training

- آموزش بدون نظارت

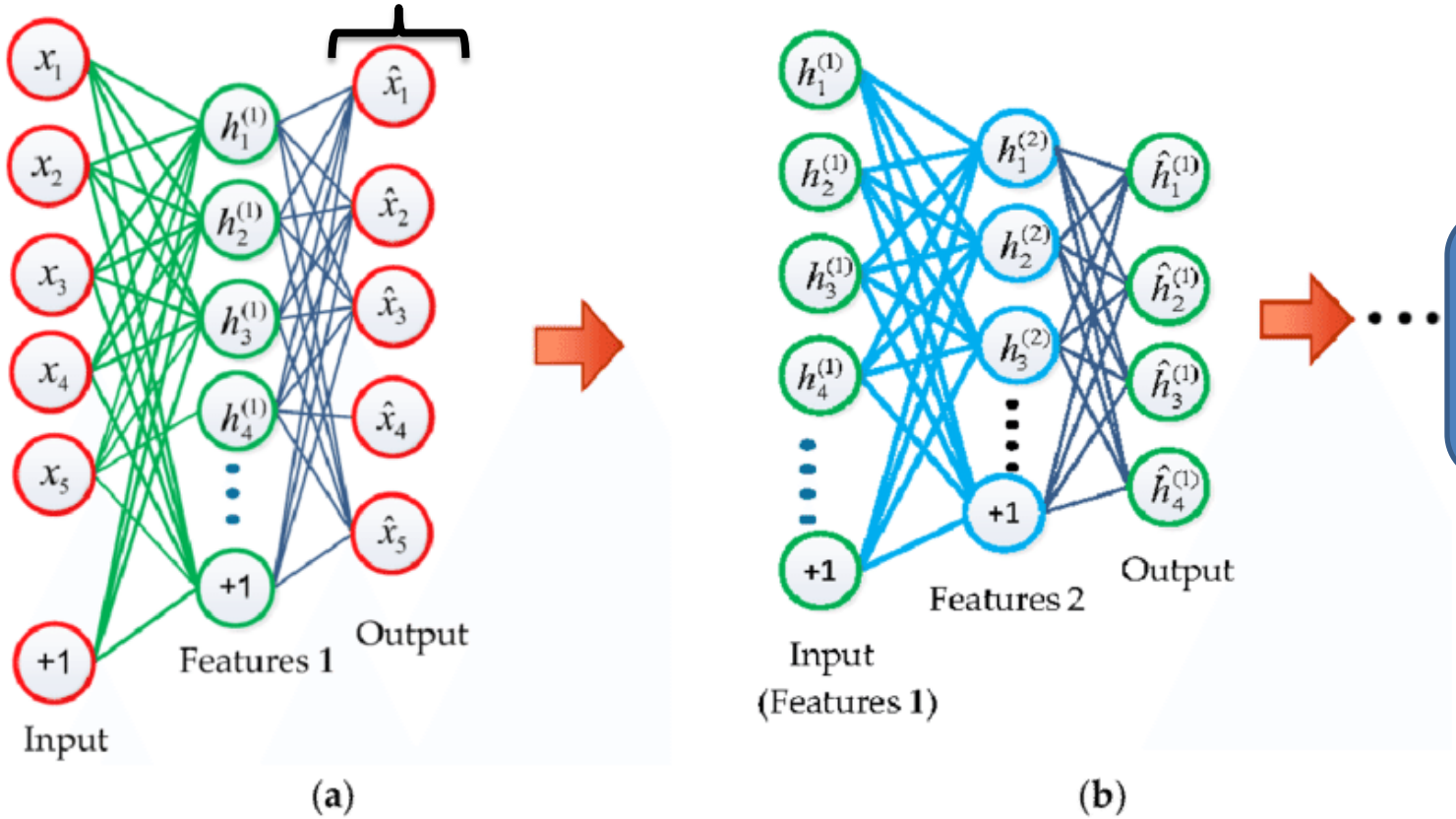


تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
گروه دایکه | dayche.com

Supervised layer-wised pre-training

- آموزش با نظارت

جایگزین با برچسب داده



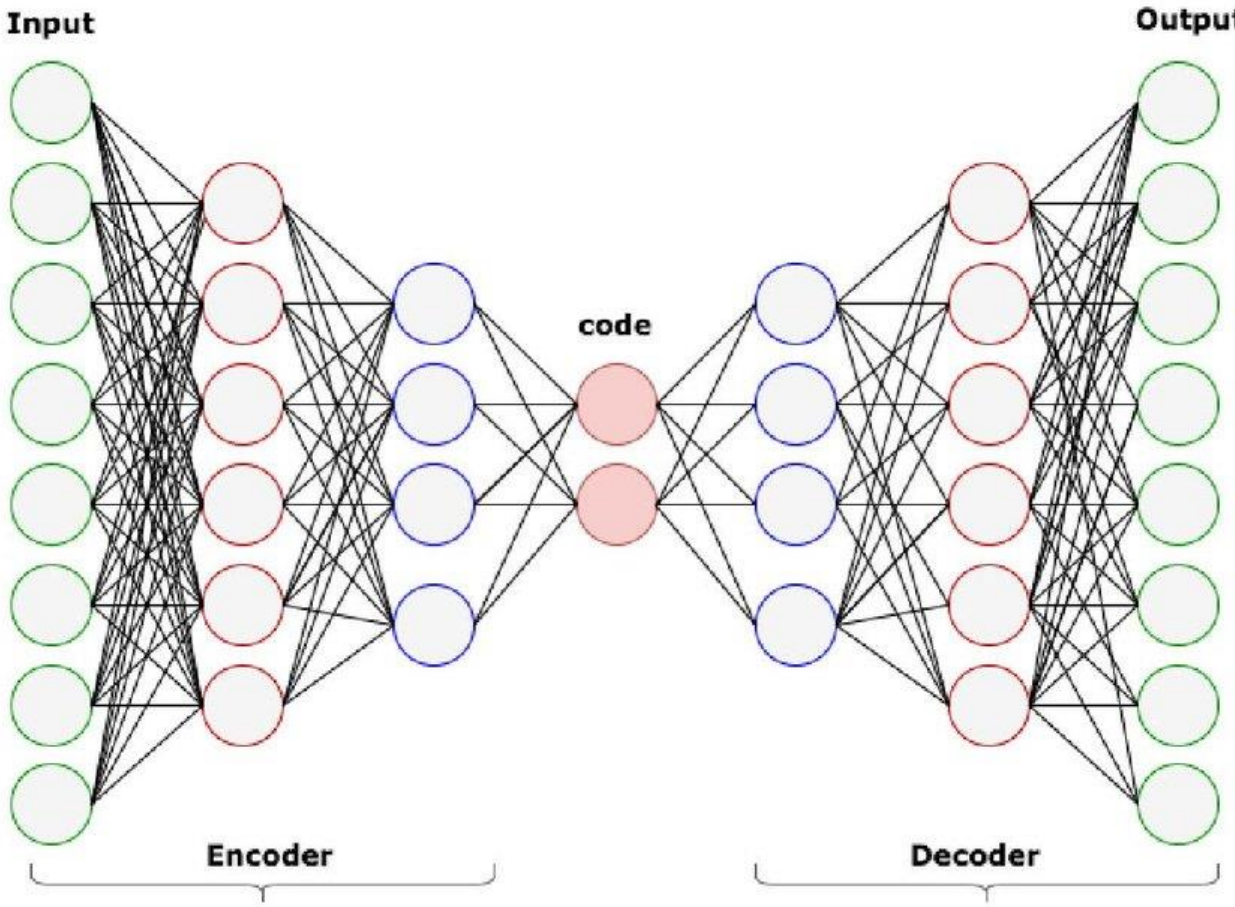
این روش به مراتب خروجی بدتری از روش آموزش بدون مربی بدست می‌دهد

- در برخی از منابع خروجی لایه اول به همراه ورودی‌ها به لایه دوم داده می‌شود

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Stacked auto-encoders

- Stacked Auto-encoders

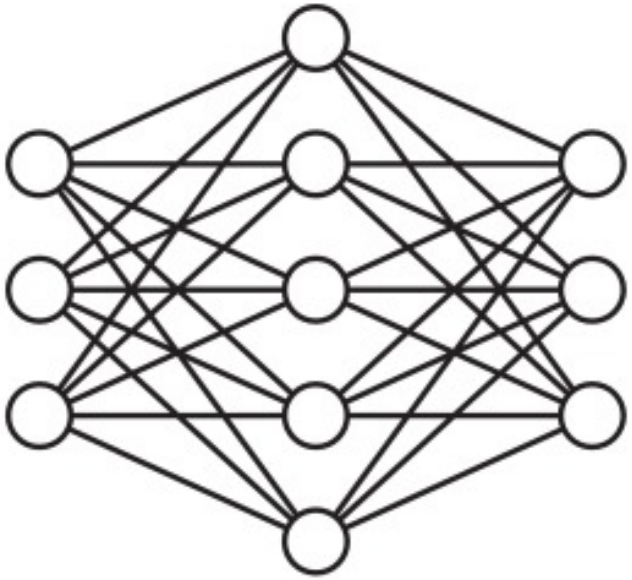


تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

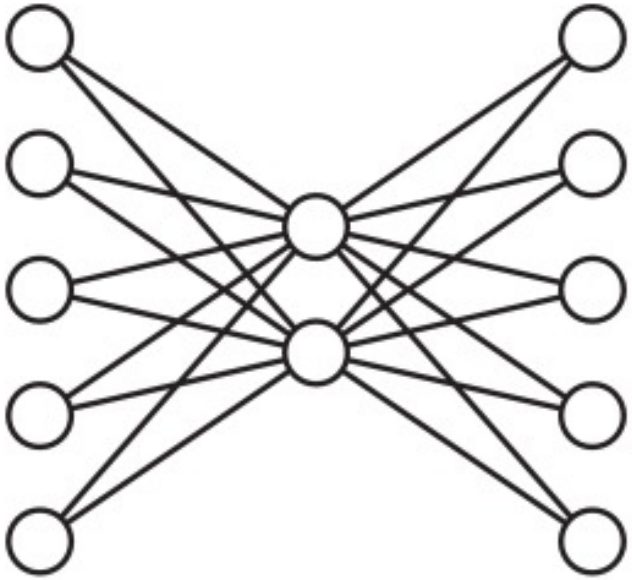
Type of auto-encoders

- انواع Auto-encoders

کدکننده فراکامل



کدکننده ناکامل

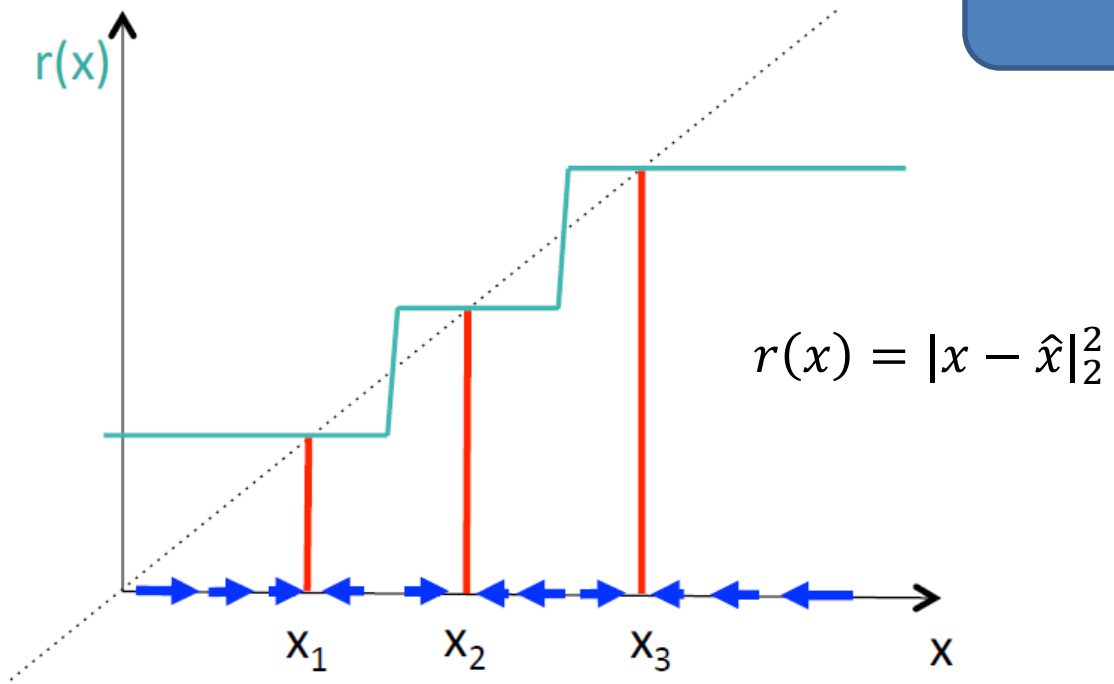


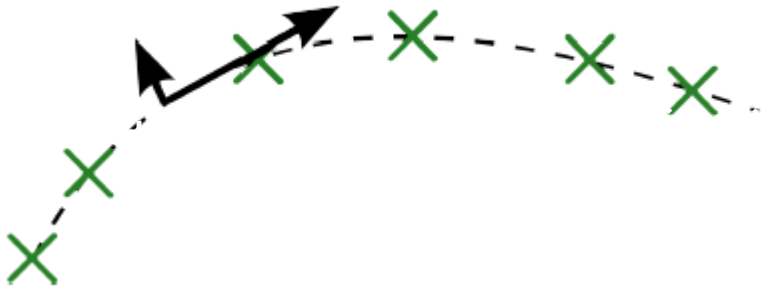
تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Overfitting

- حفظ کردن داده – مشکل از معیار بهینه‌سازی است (خطای بازیابی)

- یادگیری منیفلد: بیشتر جرم احتمال یک منیفلد در نزدیکی راستاهای بعد پایین قرار دارد.
- به منظور کاهش خطای بازیابی، نقاط نزدیک بهم خطای بازیابی متفاوتی خواهند شد، زیرا مدل سعی می‌کند خطای بازیابی را کم کند. چه اتفاقی می‌افتد؟ (فرض همواری)





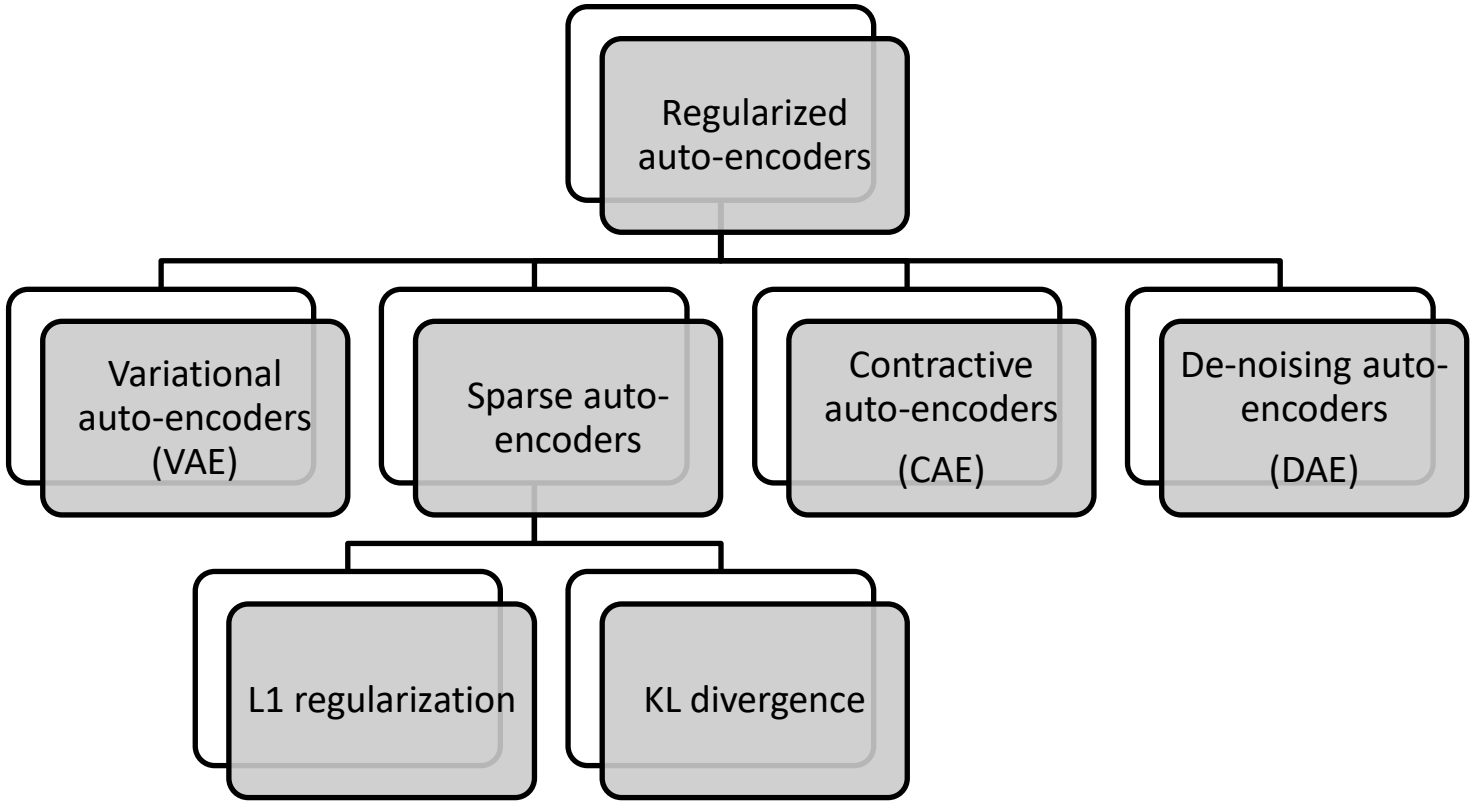
تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Regularized auto-encoders

- تنظیم کنندگی

$$L(\theta) = L(x, g(f(x)) + \phi(x, h)$$

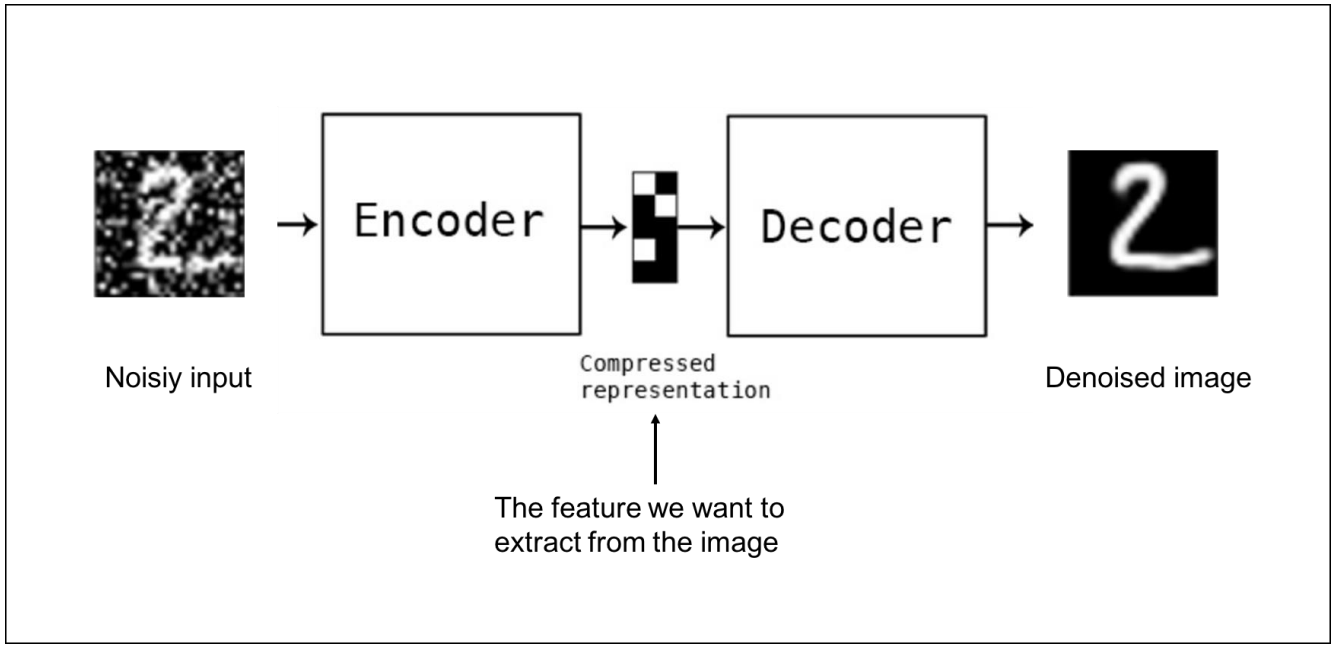
قید بر روی لایه کد



تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

De-noising auto-encoders

- کدکننده‌های خودکار نوییززدا
  - منطقا این تمایل وجود دارد که برای ورودی‌های نزدیک بهم، ویژگی‌های نزدیک بهم تولید شود.

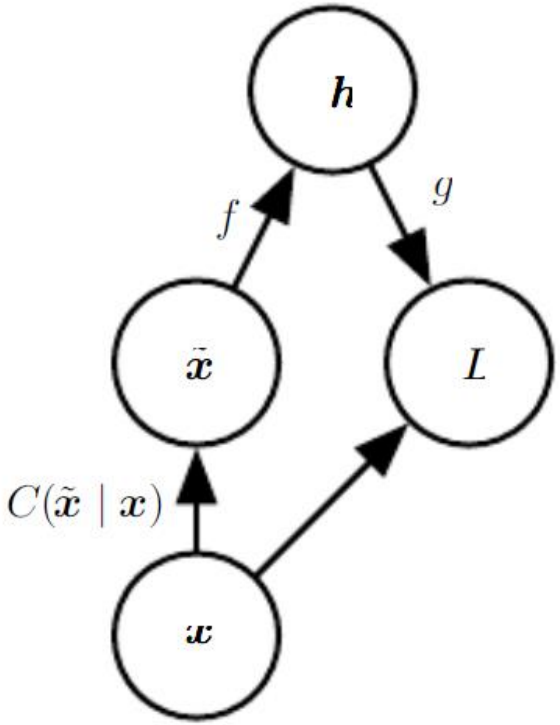


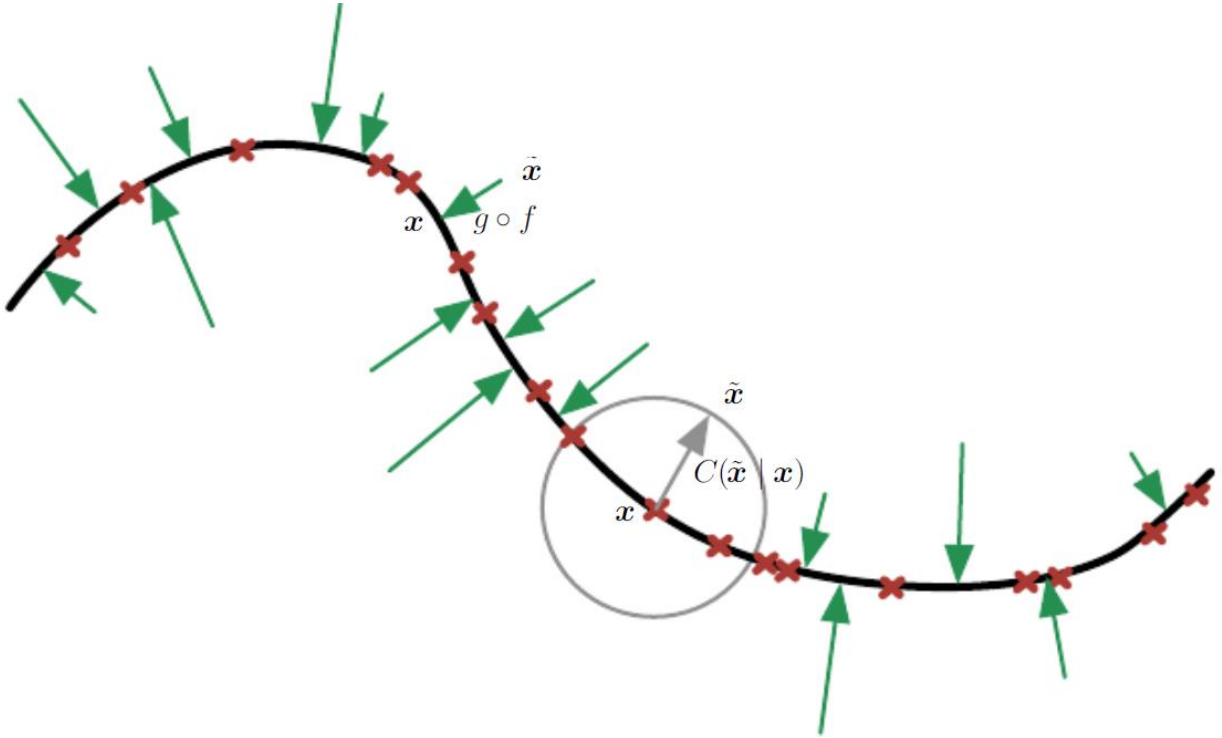
- مدل به جای یک نقطه، توزیعی از نقاط ورودی را یاد می‌گیرد.
- این مسئله قابل اثبات است که کدکننده‌های خودکار تا حدی زیادی مشخصات آماری ورودی را یاد می‌گیرند.
- این رویکرد مدل را از بیش برازش شدن دور می‌کند

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

De-noising auto-encoders

به جای یک نقطه، یک بازه از مجموع نقاط را آموزش می‌بیند



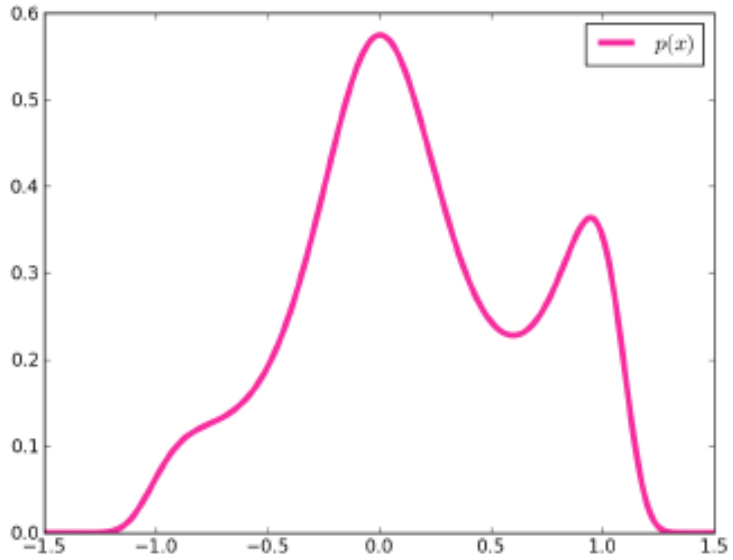


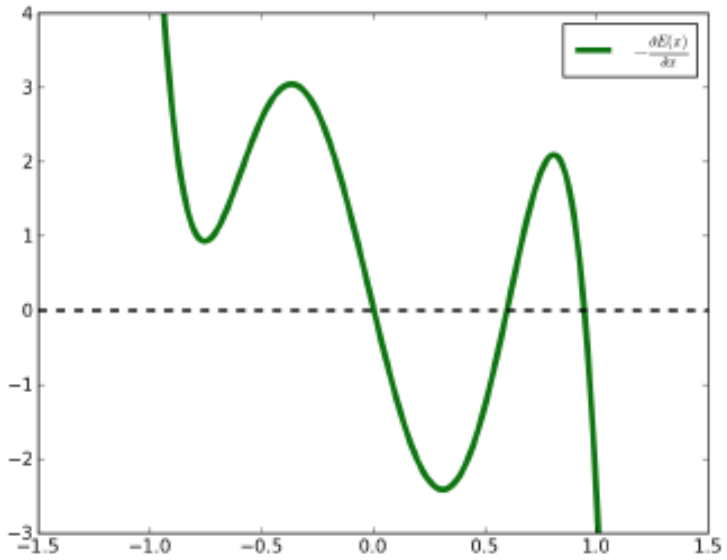
تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

De-noising auto-encoders

- کدکننده‌های خودکار نویززدا، قادر به یادگیری توزیع داده‌ها هستند. چگونه؟
  - Score matching

$$g\big(f(x)\big) - x$$





تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Contractive auto-encoders

- تابع هزینه

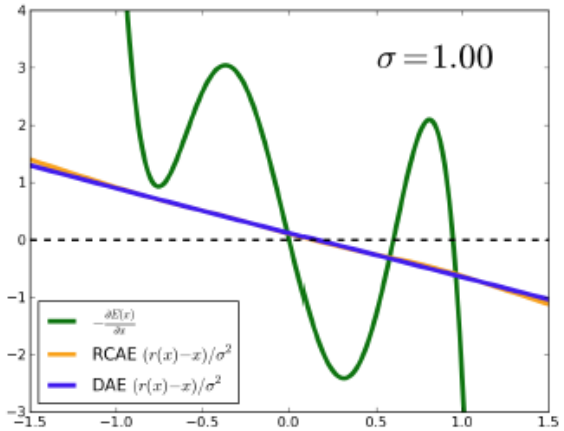
$$L(\theta) = L(x, g(f(x)) + |\nabla_x f(x)|_F^2$$

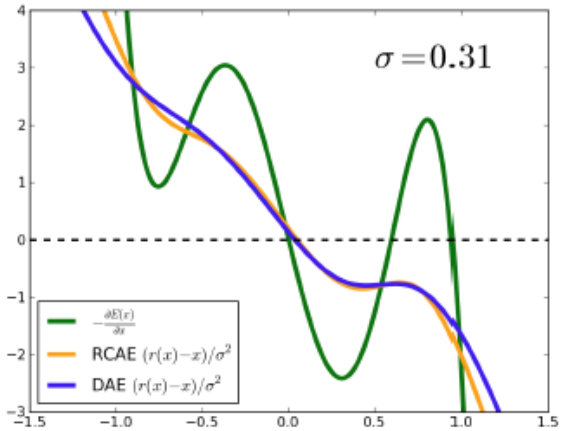
  - مطلوب است تغییرات لایه کد نسبت به تغییرات کوچک در ورودی، کوچک بماند. چه اتفاقی خواهد افتاد؟
  - آیا به طور قطع می‌توان گفت که تغییرات اعمال شده بر روی لایه کد است؟

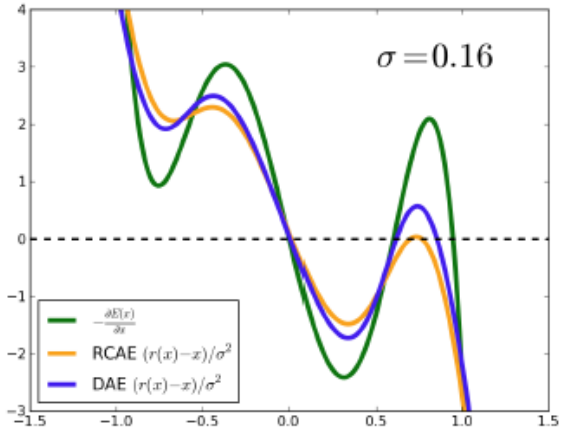
- شباهت به کدکننده نویززدار
  - نویز گوسی با میانگین صفر و واریانس کوچک اگر به ورودی افزوده شود، عملکرد دو کدکننده تقریبا یکسان خواهد بود.

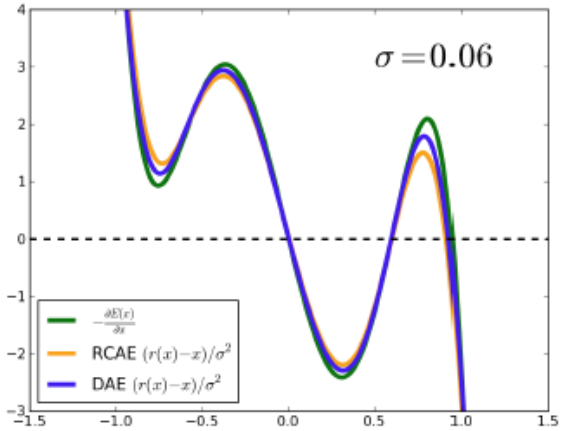
تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Contractive auto-encoders





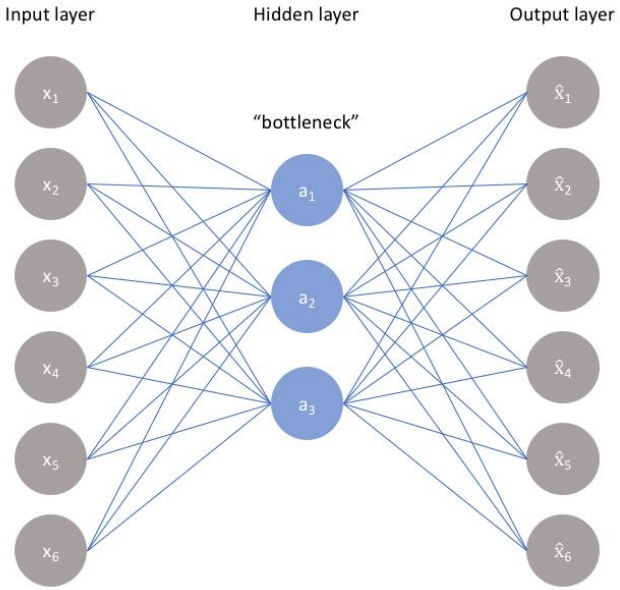




تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Sparse auto-encoders

- کدکننده‌های خودکار تنک
- فرض همواری
  - در الگوریتم‌های یادگیری ماشین، فرض همواری به صورت ضمنی وجود دارد که مطابق آن اگر $x \approx y$ باشد آنگاه $g(x) \approx g(y)$ خواهد بود



آموزش بر اساس خطای بازیابی تنظیم شده

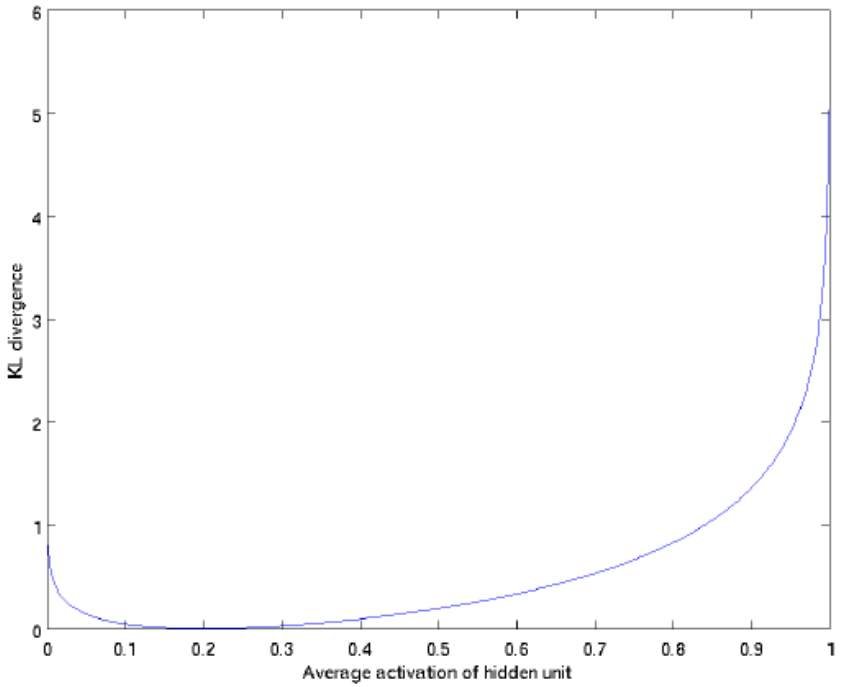
$$L(\theta) = |x - \hat{x}|_2^2 + \lambda |h|_1$$

ترم اضافه شده شرایط مشخصی را بر روی لایه کد اعمال می‌کند.
این نوع تنظیم کنندگی آیا تفسیر بیزین دارد؟
تنها وزن‌های لایه کد دستخوش تغییر می‌شوند.

Sparse auto-encoders

$$\hat{p}_j = \frac{1}{m}\sum_k \sigma_j(x_k)\,, \qquad KL\left(p\middle|\hat{p}_j\right) = p\log\frac{p}{\hat{p}_j} + (1-p)\log\frac{1-p}{1-\hat{p}_j}$$

- کدکننده‌های خودکار با معیار تنک‌زدایی KL



این تمایل وجود دارد، متوسط فعالیت هر نرون از یک توزیع مشخص پیروی کند

- فعالیت نرون به معنای روشن بودن نرون است.
- برای نرون‌های سیگموئید، متوسط فعالیت هر نرون را می‌توان به صورت میانگین فعالیت در نظر گرفت.
- برای سایر نرون‌ها چطور؟

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Sparse auto-encoders

- کدکننده‌های خودکار با معیار تنک‌زدایی KL

$$\hat{p}_j = \frac{1}{m}\sum_k \sigma_j(x_k)\,, \qquad KL\left(p\middle|\hat{p}_j\right) = p\log\frac{p}{\hat{p}_j} + (1-p)\log\frac{1-p}{1-\hat{p}_j}$$

$$J = |x - \hat{x}|_2^2 + \underbrace{\lambda \sum_{i=1}^{s} KL(p|\hat{p}_i)}$$

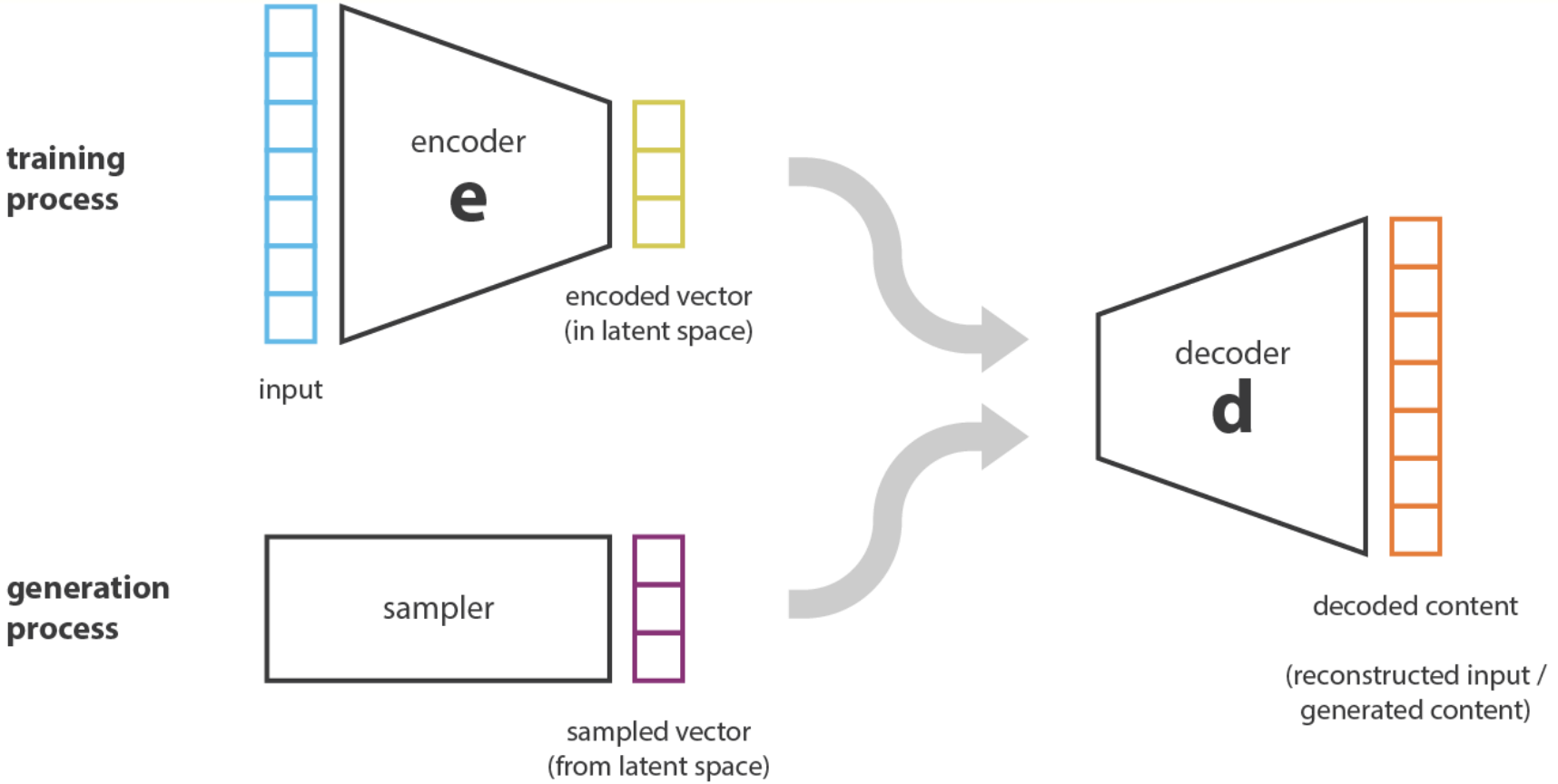
مشتق

$$\lambda\left(-\frac{p}{\hat{p}_j} + \frac{1-p}{1-\hat{p}_j}\right)$$

تنها وزن‌ها کدکننده تغییر می‌کنند.
اگر پارامترها اشتراک‌گذاری شده باشند، وزن‌های کدگشا نیز تغییر خواهند کرد.

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Variational auto-encoders



تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
گروه دایکه | dayche.com