Dimensionality reduction

Tuesday, February 20, 2024 9:33 AM

Recap from previous week

- Can fit different kinds of models to data
  - ↳ PDEs, ODEs, regressors
- Review new HW
  - focus on computation
- Discuss in class exam 3/28

This week – shifting toward dimension reduction + hidden data

Given inputs $x \in \mathbb{R}^D$, seek mapping

$$x \to z \in \mathbb{R}^L \qquad L << D$$

- Typically used for efficiency, generative modeling

  ex for images $D = \#$ pixels

- For physics, we want to reduce DOFs to make simulation orders of magnitude faster
  - → Projection-based reduced order models (PCA)
  - → Koopman operator embeddings (autoencoder + NODE)
  - → variational inference for generative models

(office hrs tomorrow)

Principal component analysis (PCA)

IDEA Postulate a linear mapping to

encode : $z = W^T x$

and decode : $\hat{x} = W z$

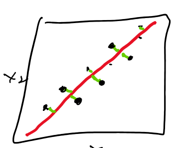
$x \triangleright z \triangleleft \hat{x}$

Seek W which minimizes a measure of reconstruction loss

$$\mathcal{L}(W) = \frac{1}{N} \sum_{i=1}^{N} \| x_i - \text{decode}_W \circ \text{encode}_W \, x_i \|^2$$

Before the math

goal identify low dimensional linear subspace



x_2 x_1

Derivation

Given dataset $\mathcal{D} = \{x_n\}_{n=1}^{N}$ scaled to have zero mean $\Rightarrow \sum_n x_n = 0$

Assume data can be decomposed into linear combination of basis functions $w_k \in \mathbb{R}^D$, $k = 1, \ldots, L$

$$x_n \approx \sum_{k=1}^{L} z_{nk} w_k$$

↑ latent "code"

Note It will turn out that w_k are the K-largest eigenvectors of covariance matrix

$$\hat{\Sigma} = Q \Lambda Q^{-1}$$

↑ first K columns

$$\mathcal{L}(W,Z) = \frac{1}{N}\|X - ZW^T\|_F^2$$
$$= \frac{1}{N}\sum_n \|x_n - Wz_n\|^2$$

Recall Orthogonal Matrix
$$Q^TQ = QQ^T = I$$

Why orthonormal? $\|x - \text{decode} \circ \text{encode}\, x\|^2$
$= \|x - W^TWx\|^2 \quad \forall x \text{ in } \text{span}(W)$
$= 0$

Strategy - proceed by mathematical induction

Recall If you want to prove $\{x_n\}_{n=1}^N$ statements are true

Principal of Mathematical Induction
① Prove x_1 is true ← Base case
② Prove x_n is true implies x_{n+1} true

will get a greedy algorithm

Base case
- Find single basis vector $w_1 \in \mathbb{R}^D$
Assuming orthonormality $w_1^Tw_1 = 1$
Let $\tilde{z}_1 = [z_{n_1}, \dots, z_{n_1}]$

$$\mathcal{L}(w_1, \tilde{z}_1) = \frac{1}{N}\sum_{n=1}^N \|x_n - z_{n_1}w_1\|^2$$
$$= \frac{1}{N}\sum_{n=1}^N (x_n - z_{n_1}w_1)^T(x_n - z_{n_1}w_1)$$
$$= \frac{1}{N}\sum_{n=1}^N x_n^Tx_n - 2z_{n_1}w_1^Tx_n + z_{n_1}^2 \underbrace{w_1^Tw_1}_{=1}$$

At optimal $\tilde{z}_1$
$$0 = \frac{\partial}{\partial z_{n_1}}\mathcal{L} = \frac{1}{N}\left[-2w_1^Tx_n + 2z_{n_1}\right]$$
$$\Rightarrow \tilde{z}_{n_1}^* = w_1^Tx_n$$

Plugging back into loss
$$\mathcal{L}(w_1, \tilde{z}_1^*) = \frac{1}{N}\sum_n x_n^Tx_n - 2z_{n_1}^2 + z_{n_1}^2$$
$$= \frac{1}{N}\sum_n \left[x_n^Tx_n - z_{n_1}^2\right]$$
$$= \text{const} - \frac{1}{N}\sum_n z_{n_1}^2$$
$$= \text{const} - \frac{1}{N}\sum_n w_1^Tx_nx_n^Tw_1$$

$$\mathcal{L}(w_1) = -w_1^T\hat{\Sigma}w_1$$
($\hat{\Sigma}$: empirical covariance)

Recall that
$$\hat{\Sigma} = \frac{1}{N}\sum_{n=1}^N (x_n - \mu)^2$$
(but we scaled data)

To find optimal unit normal basis

$$\max\ w_1^T\hat{\Sigma}w_1$$
$$\text{s.t } \|w_1\| = 1$$

Lagrange Multipliers

$$0 = \partial_{w_1}\left[w_1^T\hat{\Sigma}w_1 - \lambda_1(w_1^Tw_1 - 1)\right]$$
$$0 = 2\hat{\Sigma}w_1 - \lambda_1 2w_1$$
$$\Rightarrow \hat{\Sigma}w_1 = \lambda_1 w_1$$ eigenvalue equation

So $w^T\hat{\Sigma}w_1 = \lambda_1 w_1^Tw_1$

$= \lambda_1$ takes largest value at biggest eigenvalue

Proof by induction

Given $w_1 \dots w_{i-1}$, seek w_i satisfying

orthonormal conditions
1. $w_i^T w_j = 0$ for $j < i$
2. $w_i^T w_i = 1$

$$\mathcal{L}(w_1, \dots w_{i-1}, w_i, z_1, \dots z_{i-1}, z_i) =$$

$$= \frac{1}{N} \sum_{n=1}^{N} \| x_n - \sum_{j=1}^{i-1} z_{nj} w_j - z_{ni} w_i \|^2$$

The $\partial_{w_j} \mathcal{L} = 0$, $\partial_{z_j} \mathcal{L} = 0$ For $j < i$, we recover the minimizer from before

$z_{nj} = w_j^T x_n$

$w_j =$ j^{th} eigenvector

Fixing $j < i$ terms in loss we can show

$$0 = \frac{\partial}{\partial z_i} \mathcal{L} \implies z_{ni} = w_i^T x_n$$

$$\mathcal{L}(w_i) = \text{const} - w_i^T \hat{\Sigma} w_i$$

$$\max_{w_i} \mathcal{L}(w_i)$$

s.t $w_i^T w_i = 1$

$w_i^T w_j = 0$ for all $j < i$

$\hat{\Sigma} w_i = \lambda_i w_i$

$w_i^T \hat{\Sigma} w_i = \lambda_i$

why?

biggest is largest λ_1

but $w_i^T w_1 = 0$

next is λ_2

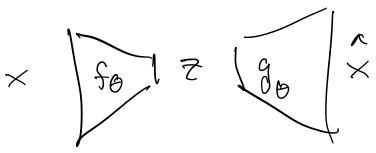
$w_i^T w_2 = 0$

etc

Guarantees

- orthonormal basis
- complete recovery

Nonlinear encodings



- Trivial to set up
- No guarantees
- No "data" for what a "good" z would be

Hidden Data

Murphy § 8.7.2

- Consider $\mathcal{D} = \{(y_i, z_i)\}_{i=1}^N$ (observable, hidden)
- log likelihood is not possible, because we have no values to plug in for z: $\log p(y, z)$ ← unknown
- Define observed log likelihood

$$LL(\theta) = \sum_{n=1}^{N} \log(y_n) = \sum_{n=1}^{N} \log\left[\sum_{z_n} p(y_n, z_n|\theta)\right]$$

- Expectation Maximization gets around this by deriving an alternative loss that bounds $LL(\theta)$ from below

Sketch

$F_\theta < LL_\theta$

$\max F_\theta < \max LL_\theta$

and if the bound is tight they are equivalent

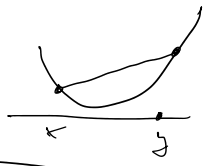
Recall Jensen's inequality

For convex $g(x)$ $\quad \mathbb{E}[g(x)] \geq g(\mathbb{E}[x])$

ex $g(x) = x^2$

formally $g(\theta x + (1-\theta)y) \le \theta g(x) + (1-\theta) g(y)$

$\forall\ \theta \in [0,1]$



Derivation

$$LL(\theta) = \sum_{n=1}^{N} \log \sum_{z_n} p(y_n, z_n \mid \theta)$$

for arbitrary dist. $q_n(z_n)$ — think of as secret dist.

$$= \sum_n \log \sum_{z_n} q_n(z_n) \frac{p(y_n, z_n \mid \theta)}{q_n(z_n)}$$

$$= \sum_n \log \mathbb{E}_{z_n \sim q}\left[\frac{p(y_n, z_n \mid \theta)}{q_n(z_n)}\right]$$

$$\ge \sum_n \mathbb{E}_{z_n \sim q}\left[\log \frac{p(y_n, z_n \mid \theta)}{q_n(z_n)}\right]$$

$$= \sum_n \sum_{z_n} q_n(z_n) \log \frac{p(y_n, z_n \mid \theta)}{q_n(z_n)}$$

$$= \sum_n \mathbb{E}_{z_n \sim q_n}\left[\log p(y_n, z_n \mid \theta)\right] + \underbrace{H(q_n)}_{\text{entropy}}$$

$$= \sum_n \underbrace{\mathcal{E}(\theta, q_n \mid y_n)}_{\text{Evidence Lower Bound (ELBO)}}$$

$H(q) = \sum_n q_n \log q_n$

We can express the ELBO

$$\mathcal{E}(\theta, q_n \mid y_n) = \sum_{z_n} q_n(z_n) \log \frac{p(y_n, z_n \mid \theta)}{q_n(z_n)}$$

$$= \sum_{z_n} q_n(z_n) \log \frac{p(z_n \mid y_n, \theta)\, p(y_n \mid \theta)}{q_n(z_n)}$$

$$\mathcal{E}(\theta, q_n \mid y_n) = \sum_{z_n} q_n \log \frac{p(z_n \mid y_n, \theta)}{q_n} - \sum q_n(z_n) \log p(y_n \mid \theta)$$

$$= -KL\left(q_n(z_n) \,\|\, p(z_n \mid y_n, \theta)\right) + \log p(y_n \mid \theta)$$

① ②

Recall KL divergence

$KL(f \| g) \ge 0$

$KL(f \| g) = 0 \Rightarrow f = g$

maximizing $\mathcal{E}$ wrt q, set $q_n^* = p(z_n | y_n, \theta)$

why? - (2) indep of q_n

(1) is least negative when $= 0$

Plug back in

$$\mathcal{E}(\theta, q^* | y_n) = \log p(y_n | \theta)$$

i.e. KL term drops out

And maximizing ELBO w/ choice

$q_n(z_n) = p(z_n | y_n, \theta)$ is tight!

↑ Posterior Dist.

end of lecture 2/20

EM - Algorithm

Given initial θ^0 params

for $t = 1, \ldots$ N iterations:

$$q_n^t = p(z_n | y_n, \theta^{t-1})$$ - E-step

$$\theta^t = \arg\max_\theta \mathbb{E}_{z_n \sim q_n^t} [\log p(y_n, z_n | \theta)]$$ - M-step

Example - Gaussian Mixture

Have data y_n but no labels for

Have data y_n but no labels for cluster ownership

$$p(y) = \sum_{c=1}^{N_{clusters}} p(y \mid z)\, p(z=c)$$

$p(z=c) \sim \text{cat}(\pi)$

$p(y \mid z) \sim \mathcal{N}(y; \mu_c, \Sigma_c)$

To derive E-step

Bayes

$$p(z=c \mid y_d) = \frac{p(y_d \mid z=c)\, p(z=c)}{p(y_d)}$$

law of total probability

$$= \frac{p(y_d \mid z=c)\, p(z=c)}{\sum_{c'} p(y_d \mid z=c')\, p(z=c')}$$

Here, interpret

$$\theta = \{\mu_c, \Sigma_c\}_{c=1}^{N_{clusters}}$$

$$q_n^t(z_n) = \frac{\pi_c\, \mathcal{N}(y_d; \mu_c, \Sigma_c)}{\sum_{c'} \pi_{c'}\, \mathcal{N}(y_d; \mu_{c'}, \Sigma_{c'})}$$

M-step

$$\mathcal{L}^t = \sum_d \mathbb{E}_{q_d^t(z_d)} \left[\log p(y_d, z_d \mid \theta) \right]$$

$$= \sum_d \sum_c q_d^t(z_d=c) \left(\log p(y_d \mid z_d, \theta)\, p(z_d \mid \theta) \right)$$

$$= \sum_d \sum_c q_d^t(z_d=c) \left[\log \mathcal{N}(y_d \mid \mu, \Sigma) + \log(\pi_c) \right]$$

$$= \sum_d \sum_c q_d^t(z_d=c)\left[\log N(y_d \mid \mu_c, \Sigma_c) + \log(\pi_c)\right]$$

Nice things

$$\frac{\partial \mathcal{L}^t}{\partial \pi_c} = 0 \Rightarrow \pi_c^{t+1} = \sum_d \frac{q_d^t(z_d=c)}{N}$$

$$\frac{\partial \mathcal{L}^t}{\partial \mu_c} = 0 \Rightarrow \mu_c^{t+1} = \frac{\sum_d q_d^t(z_d=c)\, y_d}{\sum_d q_d^t(z_d=c)}$$

$$\frac{\partial \mathcal{L}^t}{\partial \Sigma_c} = 0 \Rightarrow \Sigma_c^{t+1} = \frac{\sum q_d^t(z_d=c)\, y_d y_d^T}{\sum q_d^t(z_d=c)} - (\mu^{t+1})^T \mu^{t+1}$$

Exercise (future HW)

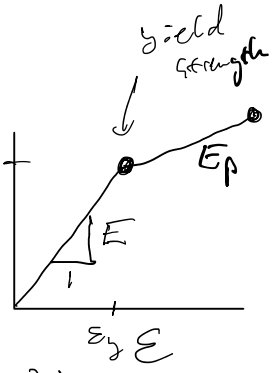
Develop a mixture of experts model which Partitions a population of N stress/strain curves into N distinct populations. Train w/ EM and use $p(z|x)$ to estimate probability a given curve belongs to a given cluster.

- Step 1

- Step 1
  - Develop expert model

For linear strain hardening model

$$\sigma(\varepsilon \mid z=c) = \begin{cases} E^c \varepsilon & \varepsilon < \varepsilon_y \\ \sigma_y^c + E_p^c(\varepsilon - \varepsilon_y^c) & \text{else} \end{cases}$$



- Define $p(y \mid z=c, \varepsilon) = N(y; \mu = \sigma(\varepsilon \mid z=c), \sigma_c^2)$
- Define $p(y, z \mid \varepsilon) = p(y \mid z=c)\, p(z=c)$ ~ cat(π)

E-step $$q_d(z_d = c) = \frac{p(y_d \mid z_d = c)\, p(z_d = c)}{\sum_{c'} p(y_d \mid z_d = c')\, p(z_d = c')}$$

M-step $$\mathcal{L} = \sum_d \sum_c q_d(z_d = c) \left[\log p(y_d \mid z_d = c, \varepsilon) + \log p(z_d = c) \right]$$

Note: During M-step, because we hold q const the derivative wrt model params $\underbrace{E^c, E_p^c, \sigma_y^c}_{\theta}$ gives

$$0 = \frac{\partial \mathcal{L}}{\partial \theta_c} = \partial_{\theta_c} \sum_c \sum_d q_d^c \log p(y \mid z=c, \varepsilon)$$

$$= \sum_c \sum_d \left[\frac{1}{\sigma^2} \sum (y_d - \mu_d^c)^2 \right]$$

$$= \partial_{\theta'} \sum_c q_d^c \sum_d \left[\frac{1}{2} \log \sigma_c^2 + \sum_\varepsilon \frac{(y_d - \mu_d^c)^2}{\sigma_c^2} \right]$$

$$0 = \partial_{\theta'} \sum_d \sum_e q_d^{c'} (y_d - \mu_d^{c'})^2$$

★ Decoupled Least squares!

Variational Auto encoders

A related idea to EM, for dim reduction

↖ "Autoencoding Variational Bayes" Kingma + Welling §20.3.5 of Murphy

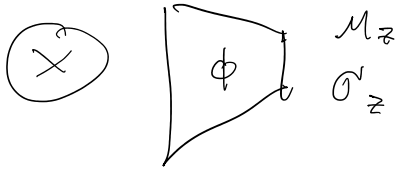
Recall Monte Carlo sampler

$$\mathbb{E}_{x \sim q}[f(x)] = \frac{1}{N} \sum_i f(x_i) \quad \text{for } x_i \sim q$$

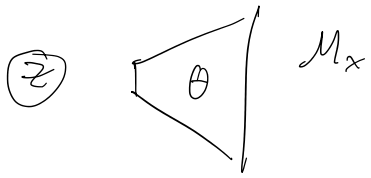
$$\lim_{n \to \infty} \left| \mathbb{E}_{x \sim q}[f(x)] - \frac{1}{N} \sum_i f(x_i) \right| = 0$$

Goal: Replace auto encoder with a probabilistic embedding

Consider $q(z|x, \phi) = \mathcal{N}(\mu_z(x), \sigma_z^2 I(x))$



$$p(x|z,\theta) = \mathcal{N}(\mu_x(z), \sigma_x^2)$$ ← constant noise



ELBO

$$\mathcal{E}(\theta,\phi|x) = \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z) - \log q_\phi(z|x)\right]$$

can manipulate

$$= \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - KL\left(q_\phi(z|x) \,||\, p(z)\right)$$

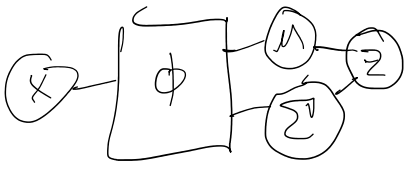
↑ sample w/ monte-carlo

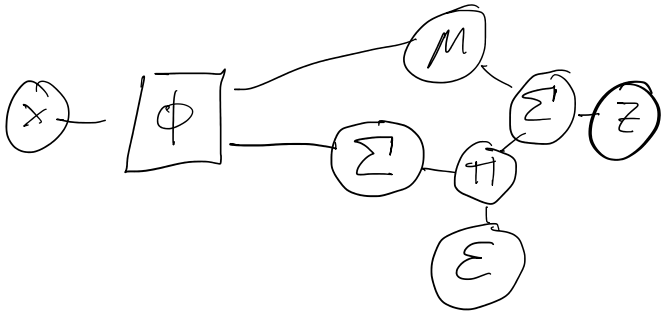
↑ a prior distribution on latent distribution

Reparameterization trick

Sample $\varepsilon \sim \mathcal{N}(0, I)$

Then $z = \mu_z(x) + \sqrt{\Sigma_x(x)}\,\varepsilon \sim q_\phi(z|x)$





Mathematically we write this

$$\mathbb{E}_{q(z|x)}\left[f(z)\right] = \mathbb{E}_{\varepsilon \sim N(0,1)}\left[f\left(\mu + \sqrt{\Sigma}\,\varepsilon\right)\right]$$