Lecture October 29

Deep Learning —
practicalities

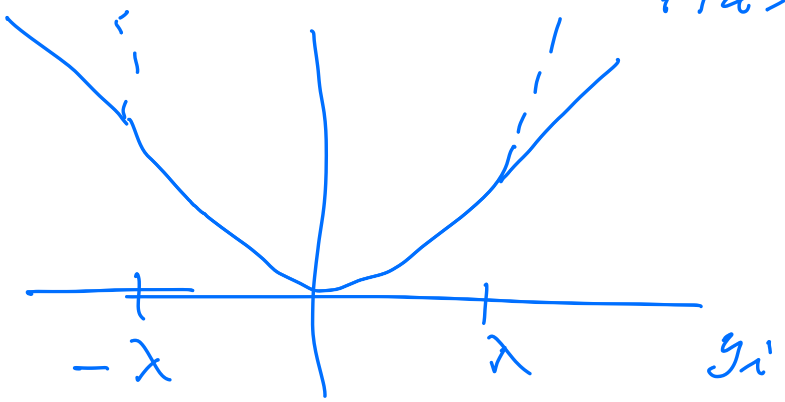
- learning rate
  - ↳ Different gradient methods
    - GD
    - SGD
    - SGD + momentum
    - Adagrad
    - RMS prop
    - ADAM
    - . . . .
- Hyperparameter λ
- architecture
  - nodes
  - layers
  - activation functions
- Cost functions

$$\frac{1}{n}\sum (y_i - \tilde{y}_i)^2 \quad \text{MSE}$$

$$\frac{1}{n}\sum (y_i - \tilde{y}_i)^2 \quad \text{when } y_i \le \lambda$$

$$+ \frac{1}{n}\sum |y_i - \tilde{y}_i| \quad \text{else}$$

Hüber



- Determine your error metric, problem driven

- Establish a working end-to-end pipeline

- Figure out bottlenecks in performance. Think of overfitting/underfitting.

- Make incremental changes
  - new data
  - adjust hyperparameter
  - method / algorithm
  -

- Most central parameter
  - learning rate η
  - hyperparameter λ

- Dimensionality - reduction
  - PCA - principal component analysis
  - Clustering

unsupervised learning

PCA

$$X \in \mathbb{R}^{n \times p}$$

$$X = \begin{bmatrix} | & | & & | \\ x_0 & x_1 & \cdots & x_{p-1} \\ | & | & & | \end{bmatrix}$$

$$n \gg p \quad (n \sim p)$$

covariance of X

$$\mathrm{cov}[X] = X^T X \frac{1}{n}$$

$$= \mathbb{E}[X^T X]$$

$$\mathrm{cov}(x_i\, x_j) = \frac{1}{n} \sum_{l=0}^{n-1} (x_{li} - \mu_{x_i}) \times (x_{lj} - \mu_{x_j})$$

$$\mathrm{var}[x_i] = \frac{1}{n} \sum_{l=0}^{n-1} (x_{li} - \mu_{x_i})^2$$

$$\mu_{x_i} = \frac{1}{n} \sum_{l=0}^{n-1} x_{li}$$

$$X = \begin{bmatrix} x_{00} & x_{01} \\ x_{10} & x_{11} \end{bmatrix} = \begin{bmatrix} | & | \\ x_0 & x_1 \\ | & | \end{bmatrix}$$

$$X^T X = \begin{bmatrix} x_{00}^2 + x_{01}^2 & x_{00}x_{01} + x_{10}x_{11} \\ x_{01}x_{00} + x_{11}x_{10} & x_{11}^2 + x_{10}^2 \end{bmatrix}$$

$$= n \begin{bmatrix} var[x_0] & cov(x_0, x_1) \\ cov(x_1, x_0) & var[x_1] \end{bmatrix}$$

PCA theorem in words:
The eigenvalues of the covariance matrix are used to reduce the number of features.

$$Cov[X] = C[X]$$

$$= \frac{1}{n} \begin{bmatrix} var[x_0] & cov(x_0, x_1) & \cdots & cov(x_0, x_p) \\ & & & \end{bmatrix}$$

$$\begin{bmatrix} \\ \mathrm{cov}(x_{p-1}, x_0) & \cdots & \mathrm{var}[x_{p-1}] \end{bmatrix}$$

orthogonal transformation

$$S \cdot S^T = S^T S = \mathbb{I}$$

$$S = \begin{bmatrix} | & | & & | \\ s_0 & s_1 & \cdots & s_{p-1} \\ | & | & & | \end{bmatrix} \in \mathbb{R}^{p \times p}$$

$$s_i^T s_j = \delta_{ij}$$

$$S^T C[x] S = \begin{bmatrix} \lambda_0 & & & \\ & \lambda_1 & & \\ & & \ddots & \\ & & & \lambda_{p-1} \end{bmatrix}$$

$$= \mathbb{E}[S^T x^T x S] = S^T C[x] S$$

$$= C[y]$$

$$= \mathrm{var}[y]$$

$$= \begin{bmatrix} \mathrm{var}[y_0] & & & \\ & \mathrm{var}[y_1] & & \\ & & \ddots & \\ & & & \mathrm{var}[y_{p-1}] \end{bmatrix}$$

Recall SVD

$$X = U \Sigma V^T \in \mathbb{R}^{n \times p}$$

$$U \in \mathbb{R}^{n \times n}$$

$$U^T U = U U^T = \mathbb{1}$$

$$V^T \in \mathbb{R}^{p \times p}$$

$$V^T V = V V^T = \mathbb{1}$$

$$\Sigma \left\{ \in \mathbb{R}^{n \times p} \right)$$

$$= \begin{bmatrix} \sigma_0 & & \\ & \ddots & \\ & & \sigma_{p-1} \\ 0 & & 0 \\ & & \vdots \\ & & 0 \end{bmatrix} = \begin{bmatrix} \tilde{\Sigma} \\ 0 \end{bmatrix}$$

$$\tilde{\Sigma} \in \mathbb{R}^{p \times p}$$

$$\tilde{\Sigma} = \begin{bmatrix} \sigma_0 & & 0 \\ & \ddots & \\ 0 & & \sigma_{p-1} \end{bmatrix}$$

$$\sigma_0 > \sigma_1 > \sigma_2 > \dots > \sigma_{p-1} > 0$$

singular values

$$X^T X = V \Sigma^T \underbrace{U^T U}_{I \in \mathbb{R}^{n \times n}} \Sigma V^T$$

$$= V \tilde{\Sigma}^2 V^T$$

$$\tilde{\Sigma}^2 = \begin{bmatrix} \sigma_0^2 & & & \\ & \sigma_1^2 & & \\ & & \ddots & \\ & & & \sigma_{p-1}^2 \end{bmatrix}$$

multiply V from the right

$$(X^T X) V = V \tilde{\Sigma}^2$$

$$V^T = \begin{bmatrix} | & | & & | \\ v_0 & v_1 & \dots & v_{p-1} \end{bmatrix}$$

$$v_i^T v_j = \delta_{ij}$$

$$(X^T X)\, v_i = v_i\, \sigma_i^2$$

$$S \cdot \quad S^T \left(X^T X \tfrac{1}{n}\right) S = \frac{1}{n} Y^T Y = \frac{1}{n} \underbrace{\begin{bmatrix} \lambda_0 & & & \\ & \lambda_1 & & \\ & & \ddots & \\ & & & \lambda_{p-1} \end{bmatrix}}_{D}$$

$$\frac{1}{n}(X^T X) S = \frac{1}{n} D \cdot S$$

$$\lambda_0 = n \cdot \sigma_0^2 \implies$$

$$\frac{1}{n}\lambda_0 = var[y_0]$$

$$\frac{1}{n}\lambda_i = var[y_i] \implies$$

The vector v_i of the SVD of X, are the eigenvectors of the covariance matrix.

And the variance is given by the singular value σ_i squared,