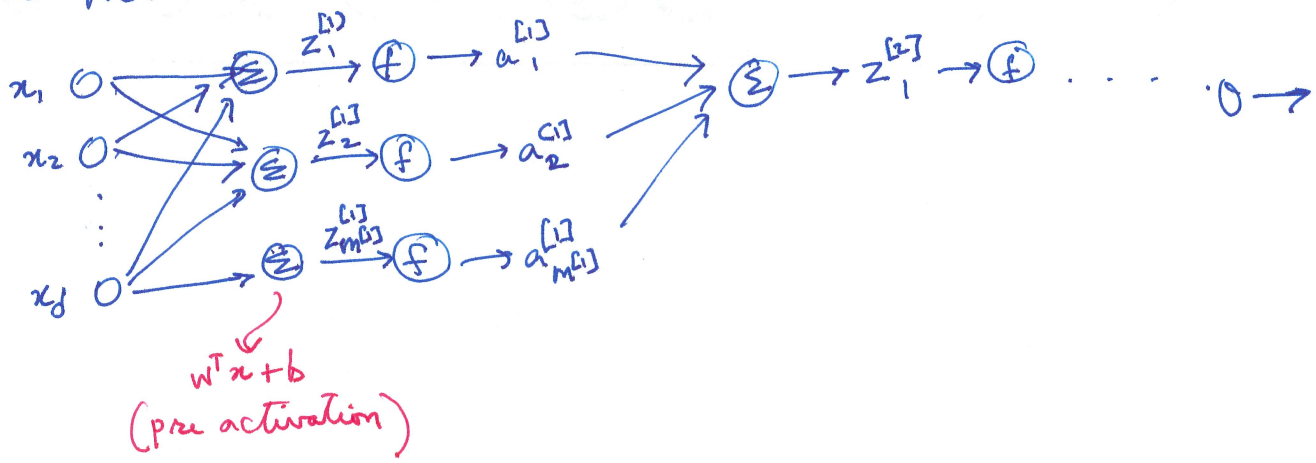
CS 307 Midterm Review

Neural Nets



Feed forward

Computing output by moving from left to right.

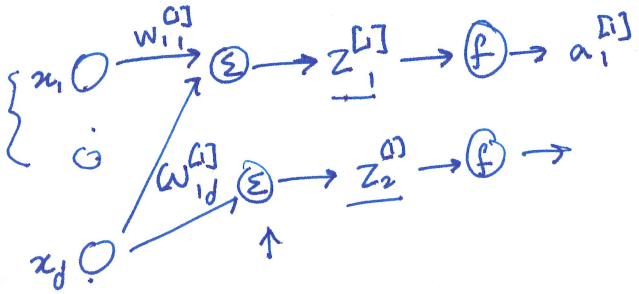
Gradients: Back propagation. Cost fn J.

$$\frac{\partial J}{\partial w_{jk}^{[i]}} = \frac{\partial J}{\partial o} \left(\frac{\partial o}{\partial w_{jk}^{[i]}} \right)$$ → Chain rule.

CNN

Filter + bias

Max pooling



Unsupervised Learning: $S = \{ x^{(1)}, x^{(2)} \ldots x^{(n)} \}$.

Reducing Dimensionality

Reducing Dimensionality:

Traing Data : d-features.

↳ Reduce → Traing : k-features.

Principal Component Analysis (PCA)

- Normalize data [mean data is 0, verriance is 1]
- Find k-directions such that projecting the training data along those k dimensions maximizes the variance of the projected data.

$S = \{x^{(1)}, .. x^{(n)}\}$.

$$X = \begin{bmatrix} - x^{(1)T} - \\ \vdots \\ - x^{(d)T} - \end{bmatrix}$$

$$A = \frac{1}{n} X^T X$$

Project X along vectors $u_1, u_2 .. u_k$ where $u_1 .. u_k$ are eigenvectors of norm 1 corresponding to the k-largest eigenvalues of A.

→ Use the SVD of X to discover the eigenvectors of A.

σ is a singular value of X iff σ^2 is eigenvalue of $X^T X$.

$$\underline{U} = \begin{bmatrix} | & | & & | \\ u_1 & u_2 & .. & u_k \\ | & | & & | \end{bmatrix}$$ — left sugular vectors of X

the columns of U are orthonormal.

→ $U^T U = I$.

$UU^T = I$ ← True.

If B is left inverse of A.

$BA = I$.

then B is a right inverse of A.

$AB = I$.

Clustering:

- Group data that is similar while keeping data that not similar in different clusters.

k-means: Find a k-partition $C_1 \dots C_k$ of S.

$C_i \cap C_j = \emptyset$ and $C_1 \cup C_2 \dots \cup C_k = S$.

s.t. ~~minimize~~

$$\min \sum_{j=1}^{k} \sum_{x^{(i)} \in C_j} \| x^{(i)} - \mu_j \|^2 \rightarrow \mu_j = \frac{\sum_{x^{(i)} \in C_j} x^{(i)}}{|C_j|}$$

k-means

pick $\mu_1 \dots \mu_k$ randomly.

Repeat.

$$y^{(i)} = \arg\min_j \| x^{(i)} - \mu_j \|^2$$

$$\mu_j = \frac{\sum_{i=1}^{n} x^{(i)} 1[y^{(i)} = j]}{\sum 1[y^{(i)} = j]}.$$

$1[b] = 1$ if b = true, 0 o.w.

Linkage-based clustering:

Start with every point being in its own cluster.

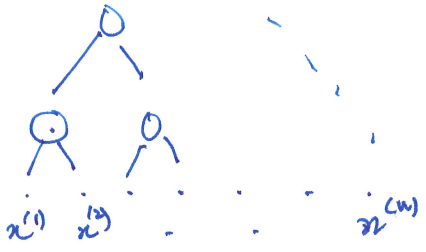
Repeat.

Merge the two clusters that are "closest".

$$dist(C, D) = \min_{x \in C, y \in D} d(x, y)$$

$$= \text{average}_{x \in C, y \in D} (d(x, y))$$

DENDROGRAM.



Training set are examples drawn from a distribution

Learning objective: Find this distribution.

→ First pick $z \in \{1 \dots k\}$. $p(z) \sim$ Multinomial

→ Pick x from a distribution that depends on z.

$p(x|z) \sim$ Gaussian distribution.
Bernoulli.

Given training set S, find parameters that define this distribution

parameters of $p(z)$
parameters of $p(x|z)$

Find parameters that maximize log-likelihood.

$$L(\theta) = \log \prod_{i=1}^{n} p(x^{(i)})$$

x, y.

$$= \sum_{i=1}^{n} \log p(x^{(i)})$$

$$= \sum_{i=1}^{n} \log \sum_{z=1}^{k} p(x^{(i)}, z)$$

$$\geq \sum_{i=1}^{n} \sum_{z=1}^{k} Q_i(z) \log \frac{p(x^{(i)}, z)}{Q_i(z)} = F(Q, \theta)$$

$Q_i(z)$ → prob distribution on z.
$\forall z \; Q_i(z) \geq 0$ & $\sum_{z=1}^{k} Q_i(z) = 1$.

Goal: Find θ that maximize $L(\theta)$.

Repeat.

$$Q^{(t+1)} = \arg\max_{Q} F(Q, \theta^t)$$

$Q_i(z) = p(z|x^{(i)})$
↳ Depends on θ^t

$$\theta^{(t+1)} = \arg\max_{\theta} F(Q^{(t+1)}, \theta)$$

→ Assumption: Can solved computationally.