CSC411H1 FALL 2018

HOMEWORK 4

LINO LASTELLA 1001237654

(1)

	# Units	# Weights	# Connections
Convolution Layer 1	290,400	34,848	105,415,200
Convolution Layer 2	186,624	614,400	111,974,400
Convolution Layer 3	64,896	884,736	149,520,384
Convolution Layer 4	64,896	1,327,104	112,140,288
Convolution Layer 5	43,264	884,736	74,760,192
Fully Connected Layer 1	4096	37,748,736	37,748,736
Fully Connected Layer 2	4096	16,777,216	16,777,216
Output Layer	1000	4,096,000	4,096,000

(b)

i. In order to reduce the number of parameters (number of weights) it would be sufficient to reduce the number of kernels in the architecture from 96 to, say, 90 at every layer.
Then, since both fully connected layers and convolution layers have number of weights $\propto$ # of kernels, the number of parameters would be reduced.

ii. For fully connected layers, the number of connections is the same as the number of weights.
For convolutional layers, the number of connections is equal to the number of weights times the size of the input (without re-counting the channels).
Then, again it would be sufficient to reduce the number of kernels in every layer to reduce the number of connections overall.

2

(d) $$P(Y=k \mid X, \mu, \sigma) = \frac{P(X \mid Y=k, \mu, \sigma) \cdot P(Y=k \mid \mu, \sigma) \cdot \cancel{P(\mu, \sigma)}}{P(X \mid \mu, \sigma) \cdot \cancel{P(\mu, \sigma)}}$$

$$= \frac{\alpha_k \cdot \left(\prod_{i=1}^{D} 2\pi\sigma_i^2\right)^{-1/2} \cdot \exp\left\{-\sum_{i=1}^{D} \frac{1}{2\sigma_i^2}(x_i - \mu_{ki})^2\right\}}{\sum_{k'} P(X \mid Y=k', \mu, \sigma)\, P(Y=k')}$$

$$= \frac{\alpha_k \cdot \left(\prod_{i=1}^{D} 2\pi\sigma_i^2\right)^{-1/2} \exp\left\{-\sum_{i=1}^{D} \frac{1}{2\sigma_i^2}(x_i - \mu_{ki})^2\right\}}{\sum_{k'}\left[\left(\prod_{i=1}^{D} 2\pi\sigma_i^2\right)^{-1/2} \exp\left\{-\sum_{i=1}^{D} \frac{1}{2\sigma_i^2}(x_i - \mu_{k'i})^2\right\} \cdot \alpha_{k'}\right]}$$

$$= \frac{\alpha_k \cdot \exp\left\{-\sum_{i=1}^{D} \frac{1}{2\sigma_i^2}(x_i - \mu_{ki})^2\right\}}{\sum_{k'} \alpha_{k'} \exp\left\{-\sum_{i=1}^{D} \frac{1}{2\sigma_i^2}(x_i - \mu_{k'i})^2\right\}}$$

□

(b) Let $D = \{ (x^{(i)}, y^{(i)}),\ i = 1, 2, \cdots, N \}$ be a dataset

with parameters $\theta = \{ \alpha, \mu, \sigma \}$.

Let $l(\theta; D) = -\log p(y^{(i)}, x^{(i)}, i = 1, 2, \cdots, N \mid \theta)$.

Then,

$$l(\theta; D) = -\log \left(\prod_{i=1}^{N} p(y^{(i)}, x^{(i)} \mid \theta) \right) \quad \text{, since data points are i.i.d.}$$

$$= -\sum_{i=1}^{N} \log p(y^{(i)}, x^{(i)} \mid \theta) \quad \text{, by basic property of logs}$$

$$= -\sum_{i=1}^{N} \log \left(p(x^{(i)} \mid y^{(i)}, \theta)\, p(y^{(i)} \mid \theta) \right)$$

$$= -\sum_{i=1}^{N} \sum_{k'} \log \left(p(x^{(i)} \mid y^{(i)} = k', \theta)\, p(y^{(i)} = k' \mid \theta) \right)$$

$$= -\sum_{i=1}^{N} \sum_{k'} \mathbb{1}\{y^{(i)} = k'\} \cdot \left[\log(\alpha_{k'}) + \log \left(\prod_{J=1}^{D} 2\pi\sigma_J^2 \right)^{-1/2} - \sum_{J=1}^{D} \frac{1}{2\sigma_J^2} \left(x_J^{(i)} - \mu_{k'J} \right)^2 \right]$$

$$= -\sum_{i=1}^{N} \sum_{k'} \mathbb{1}\{y^{(i)} = k'\} \cdot \left[\log(\alpha_{k'}) - \frac{1}{2} \sum_{J=1}^{D} \log(2\pi\sigma_J^2) - \sum_{J=1}^{D} \frac{1}{2\sigma_J^2} \left(x_J^{(i)} - \mu_{k'J} \right)^2 \right]$$

$\square$

(c)

$$\frac{\partial l(\theta;D)}{\partial \mu_{k'j}} = \frac{\partial}{\partial \mu_{k'j}} \left[-\sum_{i=1}^{N} \sum_{k'} \mathbb{1}\{y^{(i)}=k'\} \cdot \left(\log(\alpha_{k'}) - \frac{1}{2} \sum_{j=1}^{D} \log(2\pi\sigma_j^2) - \sum_{j=1}^{D} \frac{1}{2\sigma_j^2} \left(x_j^{(i)} - \mu_{k'j}\right)^2 \right) \right]$$

$$= -\sum_{i=1}^{N} \sum_{k'} \mathbb{1}\{y^{(i)}=k'\} \cdot \left(-\sum_{j=1}^{D} \frac{1}{\sigma_j^2} \left(x_j^{(i)} - \mu_{k'j}\right) \cdot (-1) \right)$$

$$= -\sum_{i=1}^{N} \cancel{\sum_{k'}} \mathbb{1}\{y^{(i)}=k'\} \cdot \cancel{\sum_{j=1}^{D}} \frac{x_j^{(i)} - \mu_{k'j}}{\sigma_j^2}$$

$$\frac{\partial l(\theta;D)}{\partial \sigma_j^2} = \frac{\partial}{\partial \sigma_j^2} \left[-\sum_{i=1}^{N} \sum_{k'} \mathbb{1}\{y^{(i)}=k'\} \cdot \left(\log(\alpha_{k'}) - \frac{1}{2} \sum_{j=1}^{D} \log(2\pi\sigma_j^2) - \sum_{j=1}^{D} \frac{(x_j^{(i)} - \mu_{k'j})^2}{2\sigma_j^2} \right) \right]$$

$$= -\sum_{i=1}^{N} \sum_{k'} \mathbb{1}\{y^{(i)}=k'\} \cdot \left(-\frac{1}{2} \sum_{j=1}^{D} \frac{1}{\sigma_j^2} - \sum_{j=1}^{D} (-1) \cdot \frac{(x_j^{(i)} - \mu_{k'j})^2}{2} \cdot \frac{1}{\sigma_j^4} \right)$$

$$= -\sum_{i=1}^{N} \sum_{k'} \mathbb{1}\{y^{(i)}=k'\} \left(-\frac{1}{2} \cancel{\sum_{j=1}^{D}} \frac{1}{\sigma_j^2} + \cancel{\sum_{j=1}^{D}} \frac{1}{2\sigma_j^4} \left(x_j^{(i)} - \mu_{k'j}\right)^2 \right)$$

To find MLE, set derivatives to zero:

$$\frac{\partial \ell}{\partial \mu_{kj}} = 0 \implies -\sum_{i=1}^{N} 1\{y^{(i)}=k\} \frac{x_j^{(i)} - \mu_{kj}}{\sigma_j^2} = 0, \quad \text{for a given } k.$$

$$\implies \mu_{kj} = \frac{\sum_{i=1}^{N} x_j^{(i)} \, 1\{y^{(i)}=k\}}{\sum_{i=1}^{N} 1\{y^{(i)}=k\}}$$

$$\frac{\partial \ell}{\partial \sigma_j^2} = 0 \implies -\sum_{i=1}^{N} \sum_{k'} 1\{y^{(i)}=k\} \left(\frac{-1}{2\sigma_j^2} + \frac{1}{2\sigma_j^4} (x_j^{(i)} - \mu_{k'j})^2 \right) = 0$$

$$\implies \frac{1}{2\cancel{\sigma_j^2}} \sum_{i=1}^{N} \sum_{k'} 1\{y^{(i)}=k'\} = \frac{1}{2\sigma_j^{\cancel{4}2}} \sum_{i=1}^{N} \sum_{k'} 1\{y^{(i)}=k'\} (x_j^{(i)} - \mu_{k'j})^2$$

$$\implies \sigma_j^2 = \frac{\sum_{i=1}^{N} \sum_{k'} 1\{y^{(i)}=k'\} (x_j^{(i)} - \mu_{k'j})^2}{\sum_{i=1}^{N} \sum_{k'} 1\{y^{(i)}=k'\}}$$

$\square$

(d) We can solve this as a maximization problem

max NLL
subject to $\sum_k \alpha_k = 1$

$$\frac{\partial l}{\partial \mu_{kj}} = 0 \rightarrow \mu_{kj} = \frac{\sum_{i=1}^{N} x_j^{(i)} \mathbb{1}\{y^{(i)}=k\}}{\sum_{i=1}^{N} \mathbb{1}\{y^{(i)}=k\}}$$ from (C)

then, by Lagrange multiplier theorem,

$$\frac{\partial l}{\partial \alpha_k} + \lambda \frac{\partial \sum_k \alpha_k}{\partial \alpha_k} = 0 \rightarrow -\sum_{i=1}^{N} \frac{\mathbb{1}\{y^{(i)}=k\}}{\alpha_k} + \lambda = 0$$

$$\rightarrow \lambda = \sum_{i=1}^{N} \frac{\mathbb{1}\{y^{(i)}=k\}}{\alpha_k}$$

From the constraint $\sum_k \alpha_k = 1$ we get $\lambda = N$ [since every class appears at least once by assumption]

Then, $\alpha_k = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{y^{(i)}=k\}$

□