Homework 4

March 5, 2019 10:12 PM

1. **[4pts] AlexNet.** For this question, you will first read the following paper:

   A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems (NIPS), 2012.

   http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks

   (a) **[3pts]** They use a conv net architecture which has five convolution layers and three fully connected layers (one of which is the output layer). Your job is to count the number of units, the number of weights, and the number of connections in each layer. I.e., you should complete the following table:

	# Units	# Weights	# Connections
Convolution Layer 1	290400	34848	105415200
Convolution Layer 2	186624	307200	223948800
Convolution Layer 3	64896	884736	149520384
Convolution Layer 4	64896	663552	112140288
Convolution Layer 5	43264	442368	74760192
Fully Connected Layer 1	4096	177209344	177209344
Fully Connected Layer 2	4096	16777216	16777216
Output Layer	1000	409600	4096000

Convolution Layer 1:

Units = 290400 (given)
Weights = 96*11*11*3 = 34,848
Connections = 55*55*96*11*11*3 = 105,415,200

Convolution Layer 2:

Units = 186624 (given)
Weights = 256*5*5*48 = 307,200
Connections = 27*27*256*5*5*48 = 223,948,800

Convolution Layer 3:

Units = 64896 (given)
Weights = 384*3*3*256 = 884,736
Connections = 13*13*384*3*3*256 = 149,520,384

Convolution Layer 4:

Units = 64896 (given)
Weights = 384*3*3*192 = 663,552
Connections = 13*13*384*3*3*192 = 112,140,288

Convolution Layer 5:

Units = 43264 (given)
Weights = 256*3*3*192 = 442,368
Connections = 13*13*256*3*3*192 = 74,760,192

Fully-Connected Layer 1:

Units = 4096 (given)
Weights = 13*13*256*2048*2 = 177209344
Connections = Weights = 177209344

Fully-Connected Layer 2:

Units = 4096 (given)
Weights = 4096 * 4096 = 16777216
Connections = Weights = 16777216

Output Layer:

Units = 1000 (given)
Weights = 4096 * 1000 = 4096000
Connections = Weights = 4096000

(b) [**1pt**] Now suppose you're working at a software company and want to use an architecture similar to AlexNet in a product. Your project manager gives you some additional instructions; for each of the following scenarios, based on your answers to Part 1, suggest a change to the architecture which will help achieve the desired objective. E.g., modify the sizes of one or more layers. (These scenarios are independent.)

i. You want to reduce the memory usage at test time so that the network can be run on a cell phone; this requires reducing the number of parameters for the network.

ii. Your network will need to make very rapid predictions at test time. You want to reduce the number of connections, since there is approximately one add-multiply operation per connection.

i. From 1. (a), we can see that the majority of weights/parameters are in the two fully connected layers. Thus one way to reduce the number of parameters for the whole network is to reduce the number Of parameters in the two fully connected layers, by reducing the size of those layers.

ii. One possible way to reduce the total number of connections is to increase the stride, which can reduce the number of kernels in a layer.

2. **[5pts] Gaussian Naïve Bayes.** In this question, you will derive the maximum likelihood estimates for Gaussian Naïve Bayes, which is just like the naïve Bayes model from lecture, except that the features are continuous, and the conditional distribution of each feature given the class is (univariate) Gaussian rather than Bernoulli. Start with the following generative model for a discrete class label $y \in (1, 2, ..., K)$ and a real valued vector of D features $\mathbf{x} = (x_1, x_2, ..., x_D)$:

$$p(y = k) = \alpha_k \tag{1}$$

$$p(\mathbf{x}|y = k, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \left(\prod_{d=1}^{D} 2\pi\sigma_d^2 \right)^{-1/2} \exp\left\{ -\sum_{d=1}^{D} \frac{1}{2\sigma_d^2}(x_d - \mu_{kd})^2 \right\} \tag{2}$$

where α_k is the prior on class k, σ_d^2 are the variances for each feature, which are shared between all classes, and μ_{kd} is the mean of the feature d conditioned on class k. We write $\boldsymbol{\alpha}$ to represent the vector with elements α_k and similarly $\boldsymbol{\sigma}$ is the vector of variances. The matrix of class means is written $\boldsymbol{\mu}$ where the kth row of $\boldsymbol{\mu}$ is the mean for class k.

(a) **[1pt]** Use Bayes' rule to derive an expression for $p(y = k|\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\sigma})$. *Hint: Use the law of total probability to derive an expression for* $p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\sigma})$.

$$p(y=k|x,\mu,\sigma) = \frac{p(y=k,x,\mu,\sigma)}{p(x,\mu,\sigma)}$$

$$= \frac{p(x|y=k,\mu,\sigma)\,p(y=k,\mu,\sigma)}{p(x|\mu,\sigma)\,p(\mu,\sigma)}$$

$$= \frac{p(x|y=k,\mu,\sigma)\,p(y=k|\mu,\sigma)\,\cancel{p(\mu,\sigma)}}{p(x|\mu,\sigma)\,\cancel{p(\mu,\sigma)}}$$

$$= \frac{p(x|y=k,\mu,\sigma)\,p(y=k|\mu,\sigma)}{p(x|\mu,\sigma)}$$

$$= \frac{p(x|y=k,\mu,\sigma)\,p(y=k)}{\sum_{j=1}^{K} p(x|y=j,\mu,\sigma)\,p(y=j)}$$

b.c. y doesn't depend on μ, σ

$$= \frac{\left(\prod_{d=1}^{D} 2\pi\sigma_d^2\right)^{-1/2} \exp\left\{-\sum_{d=1}^{D} \frac{1}{2\sigma_d^2}(x_d - \mu_{kd})^2\right\} \alpha_k}{\sum_{j=1}^{K} \left(\prod_{d=1}^{D} 2\pi\sigma_d^2\right)^{-1/2} \exp\left\{-\sum_{d=1}^{D} \frac{1}{2\sigma_d^2}(x_d - \mu_{jd})^2\right\} \alpha_j}$$

(b) **[1pt]** Write down an expression for the negative likelihood function (NLL)

$$\ell(\boldsymbol{\theta}; \mathcal{D}) = -\log p(y^{(1)}, \mathbf{x}^{(1)}, y^{(2)}, \mathbf{x}^{(2)}, \cdots, y^{(N)}, \mathbf{x}^{(N)}|\boldsymbol{\theta}) \tag{3}$$

of a particular dataset $\mathcal{D} = \{(y^{(1)}, \mathbf{x}^{(1)}), (y^{(2)}, \mathbf{x}^{(2)}), \cdots, (y^{(N)}, \mathbf{x}^{(N)})\}$ with parameters $\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\sigma}\}$. (Assume the data are i.i.d.) You may find it helpful to use the indicator notation $\mathbb{I}[y^{(n)} = k]$.

$$\ell(\theta; D) = -\log p((y^{(1)}, x^{(1)})(y^{(2)}, x^{(2)}) \cdots (y^{(N)}, x^{(N)})|\theta)$$

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

$$= -\log \prod_{j=1}^{N} p(y^{(j)}, x^{(k)} \mid \theta)$$

$$= -\log \prod_{j=1}^{N} p(y^{(j)})\, p(x^{(j)} \mid \theta)$$

$$= -\log \prod_{j=1}^{N} p(y^{(j)}) - \log \prod_{j=1}^{N} p(x^{(j)} \mid \theta)$$

$$= -\sum_{j=1}^{N} \log p(y^{(j)}) - \sum_{j=1}^{N} \log p(x^{(j)} \mid \theta)$$

$$= -\sum_{j=1}^{N} \log \left(\prod_{k=1}^{K} \alpha_k^{\mathbb{I}(y^{(j)}=k)} \right) - \sum_{j=1}^{N} \log \left(\prod_{d=1}^{D} 2\pi\sigma_d^2 \right)^{-1/2} \exp\left\{ -\sum_{d=1}^{D} \frac{1}{2\sigma_d^2} (x_d^{(j)} - \mu_{y^{(j)}d})^2 \right\}$$

$$= -\sum_{j=1}^{N} \sum_{k=1}^{K} \log \left(\alpha_k^{\mathbb{I}(y^{(j)}=k)} \right) + \frac{1}{2} \sum_{j=1}^{N} \sum_{d=1}^{D} \log(2\pi\sigma_d^2) + \sum_{j=1}^{N} \sum_{d=1}^{D} \frac{1}{2\sigma_d^2} (x_d^{(j)} - \mu_{y^{(j)}d})^2$$

$$= -\sum_{j=1}^{N} \sum_{k=1}^{K} \log(\alpha_k)\, \mathbb{I}(y^{(j)} = k) + \frac{N}{2} \sum_{d=1}^{D} \log(2\pi\sigma_d^2) + \sum_{j=1}^{N} \sum_{d=1}^{D} \frac{(x_d^{(j)} - \mu_{y^{(j)}d})^2}{2\sigma_d^2}$$

(c) **[2pts]** Take partial derivatives of the likelihood with respect to each of the parameters μ_{kd} and with respect to the shared variances σ_d^2. Based on this, find the maximum likelihood estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$. You may assume that each class appears at least once in the dataset. You may use the notation $N_k = \sum_{n=1}^{N} \mathbb{I}[y^{(n)} = k]$ in your answers.

using the answer from 2b)

$$\frac{\partial \ell(\theta; D)}{\partial \mu_{kd}} = \frac{\partial \ell}{\partial \mu_{kd}} \left(\sum_{j=1}^{N} \sum_{i=1}^{D} \frac{(x_i^{(j)} - \mu_{y^{(j)}i})^2}{2\sigma_i^2} \right)$$

$$= \sum_{j=1}^{N} \frac{\partial \ell}{\partial \mu_{kd}} \left(\sum_{i=1}^{D} \frac{(x_i^{(j)} - \mu_{y^{(j)}i})^2}{2\sigma_i^2} \right)$$

$$= -\frac{1}{\sigma_d^2} \sum_{j=1}^{N} (x_d^{(j)} - \mu_{kd})\, \mathbb{I}(y^{(j)} = k)$$

set $\frac{\partial \ell}{\partial \mu_{kd}} = 0$ to get $\hat{\mu}_{kd}$

$$\hat{\mu}_{kd} \sum_{j=1}^{N} \mathbb{I}(y^{(j)} = k) = \sum_{j=1}^{N} x_d^{(j)}\, \mathbb{I}(y^{(j)} = k)$$

$$\hat{\mu}_{kd} = \frac{\sum_{j=1}^{N} x_d^{(j)}\, \mathbb{I}(y^{(j)} = k)}{\sum_{j=1}^{N} \mathbb{I}(y^{(j)} = k)}$$

$$= \boxed{\frac{\sum_{j=1}^{N} x_d^{(j)}\, \mathbb{I}(y^{(j)} = k)}{N_k}}$$

$$\frac{\partial \ell}{\partial \sigma_d^2} = \frac{\partial}{\partial \sigma_d^2}\left(\frac{N}{2}\sum_{i=1}^{D}\log(2\pi\sigma_i^2) + \sum_{j=1}^{N}\sum_{i=1}^{D}\frac{(x_i^{(j)} - \mu_{y^{(j)}i})^2}{2\sigma_i^2}\right)$$

$$= \frac{N}{2\sigma_d^2} - \frac{1}{2\sigma_d^4}\sum_{j=1}^{N}(x_d^{(j)} - \mu_{y^{(j)}d})^2$$

set $\frac{\partial \ell}{\partial \sigma_d^2} = 0$ to find $\hat{\sigma}_d^2$

$$\frac{N}{2\hat{\sigma}_d^2} = \frac{1}{2\hat{\sigma}_d^4}\sum_{j=1}^{N}(x_d^{(j)} - \mu_{y^{(j)}d})^2$$

$$\hat{\sigma}_d^2 = \frac{\sum_{j=1}^{N}(x_d^{(j)} - \mu_{y^{(j)}d})^2}{N}$$

$$\hat{\sigma}_d = \sqrt{\frac{\sum_{j=1}^{N}(x_d^{(j)} - \mu_{y^{(j)}d})^2}{N}}$$

(d) **[1pt]** Show that the MLE for α_k is given by the following equation:

$$\alpha_k = \frac{1}{N}\sum_{n=1}^{N}\mathbb{I}[y^{(n)} = k] = \frac{N_k}{N} \tag{4}$$

You may assume that each class appears at least once. You will find it helpful to read about Lagrange multipliers[3].

$\sum_{k=1}^{K}\alpha_k = 1 \qquad \sum_{k=1}^{K}\alpha_k - 1 = 0$

using Lagrange multiplier, set $f(\theta) = \ell(\theta) - \lambda\left(\sum_{i=1}^{K}\alpha_i - 1\right)$

$$\frac{\partial f}{\partial \alpha_k} = \frac{\partial \ell}{\partial \alpha_k} - \lambda$$

$$= \frac{\sum_{j=1}^{N} I(y^{(j)} = k)}{\alpha_k} - \lambda$$

set $\frac{\partial f}{\partial \alpha_k} = 0$, we get

$$\frac{\sum_{j=1}^{N} I(y^{(j)} = k)}{\hat{\alpha}_k} = \lambda$$

$$\hat{\alpha}_k = \frac{\sum_{j=1}^{N} I(y^{(j)} = k)}{\lambda}$$

since $\sum_{k=1}^{K} \alpha_k = 1$, we have $\sum_{k=1}^{K} \sum_{j=1}^{N} I(y^{(j)} = k) = \lambda$

we know that the above expression will always be $= N$ (number of data), thus $\lambda = K$, also $\sum_{j=1}^{N} I(y^{(j)} = k) = N_k$

$$\therefore \boxed{\hat{\alpha}_k = \frac{N_k}{N}}$$