**CSC411 Homework 4**

**Question 1: AlexNet Analysis**

**1a: Counting Parameters**

Formulas for Analysis

Convolutional (Conv) Layers

Dense Layers

|  |  |  |  |
| --- | --- | --- | --- |
| **Layer** | **Units** | **Weights** | **Connections** |
| **Conv 1** |  |  |  |
| **Conv 2** |  |  |  |
| **Conv 3** |  |  |  |
| **Conv 4** |  |  |  |
| **Conv 5** |  |  |  |
| **Full 1 (Dense)** |  | \* | |
| **Full 2 (Dense)** |  |  | |
| **Output (Dense)** |  |  | |

\*I’m ignoring max pooling that should take place between Conv 5 and Full 1: <https://piazza.com/class/jlp72odwmqo2v2?cid=606>

**1b: Case Study**

1. Cell Phone: Parameter / Weight Reduction
   1. Decrease number of dense units. For example, reducing the number of dense units to 2048 in Full 1 would cut the number of parameters by . This would reduce the representational capacity of the dense layers but drastically decrease the amount of weights.
2. Rapid Predictions: Connection Reduction
   1. Decrease dimensionality of convolution layer outputs. Can make smaller output tensors by:
      1. Reducing number of kernels (means less output channels)
      2. Increasing stride of kernels (means smaller output images, and less dot-products per image)
      3. Increasing max-pooling patch size (means less output pixels)
   2. Decrease number of dense units (parameter reduction mentioned above also carries over to number of computations / connections)

**Question 2: Naïve Bayes Analysis**

**Denote for brevity:**

Formula for Reference

**Part A**

Applying Bayes Theorem:

Total probability can be applied to yield

is a prior, can simplify conditionals involving it

Rephrasing equation

**Part B**

Note: M refers to number of dimensions as D now refers to data

**Part C**

Derivative WRT (second and third terms are only ones that depend on )

MLE for : biased variance of attribute over ALL samples. Proof:

Derivative WRT (only term that needs to be considered is the third one)

(feature) is fixed if we are computing in terms of – can drop the outer sum for simplicity

Use chain rule to compute derivative, and binary equality function to “turn off” terms whose class isn’t k (their means are different):

MLE for (mean for j-th attribute for class k) – average of attribute j, for all samples from class k. Proof:

**Part D**

I was unable to complete this question.