## **Computing derivatives w.r.t Output Layer**

### **Part 3**

1. So far, we have derived the partial derivative with respect to the *i*-th element of layer aL
2. We can now write the gradient w.r.t the vector aL
3. As we saw earlier, aL = [aL1, aL2 … aLk]
4. Going by the indicator variable in step 1, it resolves to 0 for all values of *i* except for *i* = *l*
5. Let us assume a scenario where k = 4, and *l* = 2
   1. Here, the indicator variable values as a vector would be
6. The gradient w.r.t a*L* is

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| .  .  . |
|  |
|  |  |  |  |
| .  .  . |
|  |

1. The above can be seen as a difference of two vectors, [0, 1, 0, … 0k] and ŷ
2. The first vector is essentially the one hot representation of the true output e(*l*):
3. In reality, this is simply the difference between the true distribution y and the predicted distribution ŷ