Aaron Dardik

CS567 Problem Set #2

* 1. ) Given we will denote by We will let Now we must show that is in the exponential family. Now, By using the fact that the logarithm of a product is the sum of the logarithms the above expression simplifies to Let (written here using the variable for convenience instead of This shows that the categorical distribution is in the exponential family.
  2. ) Bernoulli This simplifies to Now, Now,
  3. ) From part 1.1 we see that for the categorical distribution we have . From this we see that By substituting back in and dividing we see that Now, as Now, we note that can be expressed as a linear combination of so we have that with being a parameter expressed in terms of the first and not an independent variable itself.
  4. ) From this we conclude that the Poisson distribution is in the exponential family with Now, As the expected value of a Poisson distribution is we have

2.1.1 ) For ease of notation we will denote our cost function by instead of To find it suffices to find We will consider two cases: In the first case, we have The latter derivative we will solve via the quotient rule to get Combining these results we have In the latter case where we have Thus, letting be the vector whose th coordinate is 1 if and 0 otherwise, we have that

2.1.2 ) Now, Additionally,

2.1.3 ) where

2.1.4 ) Additionally, we have

2.2 ) being initialized to zero implies that and is constant. Now, plugging these values in we have Now, learning in the hidden layer would adjust along the gradient of cost with respect to those variables, but as the partial derivatives along those variables are zero, the gradient is zero and no learning occurs.

2.3 ) Removing the nonlinear operation we have Letting Let and and is therefore linear in

3.1 ) Taking derivatives and setting equal to zero, we have Now, We now plug in, to get the update rule . From this we can simplify to the update rule

3.2 ) If we attempt to do gradient descent on regularized linear regression without kernel, then we have to recalculate for all each time.

3.3.1 ) Note If we see that is a linear combination of Assume then that the inductive hypothesis holds for time step up to Now, as we can substitute back into the expression to get = which is a linear combination of and so the result holds.

3.3.2 ) Substituting for we get

4.1 ) Taking the derivative of the expression “inside” the arg max we have

4.2 ) By multiplying both sides of the equation by and simplifying notation, we have Dividing by 4 and remembering that we know the square root of the right expression (as we calculated it by squaring another expression) we can conclude the following result: .

4.3 ) We cannot always solve a problem in terms of its dual formulation, if strong duality does not hold then the that minimizes training error may not be the same as that which maximizes