Lecture 2 optional exercise: training a critic

D.J. Wischik

Let p be the supposed true distribution from which MNIST digits are drawn, and let q be the restriction to even digits. Make yourself a bag of samples from each, using disjoint images, and split each bag into training, validation and holdout sets.

1. Train a classifier $d(x) \approx \Pr(\text{from } q \mid x)$ to tell the bags apart, and let $\lambda(x) = d(x)/(1-d(x))$.
2. Use holdout data to estimate

$$L(q, p, \lambda) := 1 + \mathbb{E}_{X\sim q} \log \lambda(X) - \mathbb{E}_{X\sim p}\lambda(X).$$

   What does this value tell you? Answer with maths, not words.

3. Overtrain your classifier. What is L now? Comment.
4. Write $\hat{q}^{\text{tr}}$ and $\hat{q}^{\text{ho}}$ for the empirical distributions of the training and holdout bags from q, and likewise for p. Let λ be some classifier fitted on the training sets. Discuss the difference between $L^{\text{tr}} := L(\hat{q}^{\text{tr}}, \hat{p}^{\text{tr}}, \lambda)$ and $L^{\text{ho}} := L(\hat{q}^{\text{ho}}, \hat{p}^{\text{ho}}, \lambda)$. If λ is overtrained, what happens to L^{tr} and L^{ho}?
5. What is $\mathrm{KL}(q \,\|\, p)$ for these specific datasets?
6. State a bound on $\mathrm{KL}(p \,\|\, q)$ that uses your classifier from step 1. What is $\mathrm{KL}(p \,\|\, q)$ for these specific datasets?