CSE 546 HW #2

Sam Kowash

November 1, 2018

(1) A Taste of Learning Theory

1. Let $X \in \mathbb{R}^d$ a random feature vector, and $Y \in \{1, \ldots, K\}$ a random label for $K \in \mathbb{N}$ with joint distribution P_{XY}. We consider a randomized classifier $\delta(x)$ which maps a value $x \in \mathbb{R}^d$ to some $y \in \{1, \ldots, K\}$ with probability $\alpha(x, y) \equiv P(\delta(x) = y)$ subject to $\sum_{y=1}^{K} \alpha(x, y) = 1$ for all x. The risk of the classifier δ is

   $$R(\delta) \equiv \mathbb{E}_{XY,\delta}\left[\mathbf{1}\{\delta(X) \neq Y\}\right],$$

   which we should interpret as the expected rate of misclassification. A classifier δ is called deterministic if $\alpha(x, y) \in \{0, 1\}$ for all x, y. Further, we call a classifier δ_* a Bayes classifier if $\delta_* \in \arg\inf_\delta R(\delta)$.

   If we first take the expectation over outcomes of δ (by conditioning on X and Y), we find

   $$R(\delta) = \mathbb{E}_{XY}\left[1 - \alpha(X, Y)\right],$$

   since the indicator function is 1 except for the single outcome where $\delta(x) = y$, which occurs with probability $\alpha(x, y)$. It is then clear that minimizing $R(\delta)$ is equivalent to *maximizing* $\mathbb{E}_{XY}[\alpha(X, Y)]$; the assignments of $\alpha(x, y)$ which do this are our Bayes optimal classifiers.

2. We grab n data samples (x_i, y_i) i.i.d. from P_{XY} where $y_i \in \{-1, 1\}$ and $x_i \in \mathcal{X}$ where $\mathcal{X}$ is some set about which we make no further assumptions.

(2) Programming

1.

2.

3. We now consider binary classification between 2s and 7s in the MNIST set via regularized logistic regression. We choose a balanced target set $Y \in \{-1, 1\}$, where $Y = -1$ for 2s and $Y = 1$ for 7s, so that our data are $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{Z}_2$. The L_2-regularized negative log likelihood objective to be minimized is

   $$J(w, b) = \frac{1}{n}\sum_{i=1}^{n} \log\left[1 + \exp\left(-y_i(b + x_i^T w)\right)\right] + \lambda\|w\|_2^2.$$

   For convenience, we define the functions

   $$\mu_i(w, b) = \frac{1}{1 + \exp\left[-y_i(b + x_i^T + w)\right]}.$$

   .

(a) To do gradient descent, we need to know some gradients. First,

$$
\begin{aligned}
\nabla_w J(w,b) &= \frac{1}{n}\sum_{i=1}^{n} \frac{-y_i x_i \exp\left[-y_i(b + x_i^T w)\right]}{1+\exp\left[-y_i(b+x_i^T w)\right]} + 2\lambda w \\
\nabla_w J(w,b) &= -\frac{1}{n}\sum_{i=1}^{n} \mu_i \left(\frac{1}{\mu_i} - 1\right) y_i x_i + 2\lambda w \\
\nabla_w J(w,b) &= \frac{1}{n}\sum_{i=1}^{n} (\mu_i - 1) y_i x_i + 2\lambda w.
\end{aligned}
$$

Next,

$$
\begin{aligned}
\nabla_b J(w,b) &= -\frac{1}{n}\sum_{i=1}^{n} \frac{y_i \exp\left[-y_i(b + x_i^T w)\right]}{1+\exp\left[-y_i(b+x_i^T w)\right]} \\
\nabla_b J(w,b) &= \frac{1}{n}\sum_{i=1}^{n} (\mu_i - 1) y_i.
\end{aligned}
$$