Homework 3

Xixiang Chen

1 Separability

Let x_n, x'_m be two sets of linearly separable points, and (w, w_0) the separator s.t.

$$w^T x_n + w_0 > 0 \tag{1}$$
$$w^T x'_m + w_0 < 0 \tag{2}$$

From (1) - (2) we have:

$$w^T(x_n - x'_m) > 0 \tag{3}$$
$$w^T(\sum_n \alpha_n x_n - \sum_m \beta_m x'_m) > 0 \tag{4}$$
$$so\ \sum_n \alpha_n x_n - \sum_m \beta_m x'_m \neq 0\ for\ all\ \alpha_n, \beta_m \tag{5}$$

Assume their convex hulls intersect, there must be a point in both $\mathbf{x}$ and $\mathbf{x}'$, which means there exists a pair of (α_n, β_m) s.t.

$$\sum_n \alpha_n x_n = \sum_m \beta_m x'_m$$

This is a contradiction with equation (5). Therefore their convex hulls must not intersect.

2 Logistic Regression and Gradient Descent

(a)

$$\begin{aligned}
\sigma'(a) &= \frac{d}{da}\frac{1}{1+e^{-a}} \\
&= \frac{d}{da}(1+e^{-a})^{-1} \\
&= -1 \cdot (1+e^{-a})^{-2} \cdot e^{-a} \cdot -1 \\
&= \frac{e^{-a}}{(1+e^{-a})^2}
\end{aligned}$$

$$\begin{aligned}
\sigma(a)(1-\sigma(a)) &= \frac{1}{1+e^{-a}} \cdot (1 - \frac{1}{1+e^{-a}}) \\
&= \frac{1}{1+e^{-a}} \cdot \frac{e^{-a}}{1+e^{-a}} \\
&= \frac{e^{-a}}{(1+e^{-a})^2}
\end{aligned}$$

$$Therefore\ \sigma'(a) = \sigma(a)(1-\sigma(a))$$

(b)

Using the result from the previous question:

$$\begin{aligned}
\frac{\partial L_w(\{(x_i), y_i\}_{i=1}^n)}{\partial w_j} &= \sum_{i=1}^{n} -y_i \cdot \frac{h_w(x_i)(1-h_w(x_i))}{h_w(x_i)} \cdot x_{ij} - (1-y_i) \cdot \frac{1}{1-h_w(x_i)} \cdot (-1) \cdot h_w(x_i)(1-h_w(x_i)) \cdot x_{ij} \\
&= \sum_{i=1}^{n} -y_i x_{ij} + y_i h_w(x_i) x_{ij} + h_w(x_i) x_{ij} - y_i h_w(x_i) x_{ij} \\
&= \sum_{i=1}^{n} -y_i x_{ij} + h_w(x_i) x_{ij}
\end{aligned}$$

(c)

To prove that the cross entropy loss function is convex, we show that its hessian matrix with respect to **w** is positive semi-definite:

$$\begin{aligned}
\nabla_w^2\{-ylog[h_w(x)] - (1-y)log[1-h_w(x)]\} &= \nabla_w(\nabla_w\{-ylog[h_w(x)] - (1-y)log[1-h_w(x)]\})\} \\
&= \nabla_w[-yx + h_w(x)x] \\
&= xh_w(x)(1-h_w(x))x \\
&= h_w(x)(1-h_w(x))xx^T
\end{aligned}$$

For any vector **v**:

$$\begin{aligned}
v^T h_w(x)(1-h_w(x))xx^T v &= h_w(x)(1-h_w(x))v^T xx^T v \\
&= h_w(x)(1-h_w(x))(x^T v)^2 \\
&\geq 0
\end{aligned}$$

By definition this hessian matrix is positive semi-definite. Since the sum of two (or more) convex functions (for all x_i) is also convex, we conclude that the cross entropy loss function is convex.

(d)

- How many reviews were predicted to have high rating?
- 26998

- What is the accuracy of the model on predictions made above? (round to 2 digits of accuracy)
- 0.66

- What are the top 3 most positively weighted words (according to our model)?
- love, loves, easy

3 Boosting

(a)

$$Miscl.\ error = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{[y_i \neq f(x_i)]} \leq R^{train} \leq e^{-2\gamma_{WLA}^2 T}$$

As T increases, the RHS goes to 0. Therefore the misclassification error equals zero eventually.

(b)

Assume weight vector $\boldsymbol{w}$ for the data point is normalized. Then the weighted version is setting the initial weight of data point x_i to $\frac{w_i}{n}$, and the rest is the same.

(c)

- Are the weights monotonically decreasing, monotonically increasing, or neither?
- Neither.

- From this plot (with 30 trees), is there massive overfitting as the of iterations increases?
- No. As illustrated in the figure, test error is below training error.