CSE 547 ML for Big Data

Homework 1

Problem 1.1

(a) **False.** When x_k^j is linearly separable, the weights approach infinity and therefore could lead to multiple solutions. This possibility is eliminated by the fact that regularization with $\lambda > 0$ penalizes large weights, and there remains one globally optimal solution.

(b) **False.** The l_2 norm reduces the feature space but typically does not force $\hat{\mathbf{w}}$ to be very sparse. This could be compared to l_1 normalization, in which many of the values of $\hat{\mathbf{w}}$ become zero and the feature space becomes very sparse.

(c) **True.** When the data is linearly separable, the maximum likelihood increases as the weights increase. Thus, with no regularization (i.e. $\lambda = 0$), the weights will approach infinity, and the sigmoid function becomes more like a step function.

(d) **True.** As the regularization term overpowers $l(\hat{\mathbf{w}}, \mathcal{D}_{\text{train}})$, lower weights are favored and the loss increases.

(e) **False.** When the regularization term increases, the loss can also decrease for unseen data because the test set's performance benefits from generalization.

(f) **True.** Let's say we have a misclassifed point on one side of the decision boundary. If we duplicate that point, the function has a stronger incentive to change the decision boundary to incorporate that point correctly.

(g) i.

$$L(w_0, w_1) = \frac{1}{n}\sum_i log(P(Y_i|x_i, w) = \frac{1}{1 + exp(w_0 + w_1 x_1)}$$

$$\begin{aligned} L(w_0', w_1', w_2') &= \frac{1}{n}\sum_i log(P(Y_i|x_i, w) = \frac{1}{1 + exp(w_0' + w_1' x_1 + w_2' x_1)} \\ &= \frac{1}{n}\sum_i log(P(Y_i|x_i, w) = \frac{1}{1 + exp(w_0' + w_1' x_1 + w_2' x_1)} \end{aligned}$$

ii. w_0 and w_0' should be the same.
$w_1 = w_1' + w_2'$.

This linear combination of weights w_1' and w_2' increases the possibility of achieving a different decision boundary than if we had just one of the weights. For instance, in an extreme case, $w_1' = 0$, and $w_2' = w_1$. In this case (without regularization) we could potentially change the decision boundary if given a duplicate variable.

iii. Yes. For instance, w_1' and w_2' would converge upon the same values because L2 normalization favors smaller weights. This would then cause the decision boundary to remain the same.

Problem 1.2

(a) We need to estimate R-1 parameters (not including the pseudoparameter vector w_R – it's all 0's. These parameters each represent a set of weights for the logistic regression for that class.

(b) for $k < R$:

$$\begin{aligned} L(w_1, ..., w_{R-1}) &= \sum_{j=1}^{N} ln(P(y^j|x^j, w)) \\ &= \sum ln\left(\frac{exp(w_k^T x^j}{1 + \sum_i exp(w_j^T x^j)}\right) \\ &= \sum_j (ln(exp(w_k^T x^j) - ln(1 + \sum_i exp(w_i^T, x^i)) \\ &= \sum_j (w_k^T x^j - ln(1 + \sum_i exp(w_i^T, x^i)) \end{aligned}$$

(c)

$$\mathbf{\Delta L} = \sum_j \left(x^j - \frac{exp(w_k^T x^j)x^j}{1 + \sum_i exp(w_j^T x^j)} \right)$$

(d)

$$\begin{aligned} L(w_1, ..., w_{R-1}) &= \sum_{j=1}^{N} lnP(y^j|x^j, w) - \frac{\lambda}{2} \sum_{l=1}^{R-1} ||w_l||_2^2 \\ &= \sum_{j=1}^{N} \left(w_k^T j^j - ln(1 + \sum_i exp(w_i^T x^i)) \right) - \frac{\lambda}{2} \sum_{l=1}^{R-1} ||w_l||_2^2 \end{aligned}$$

$$\mathbf{\Delta L} = \sum_j \left(x^j - \frac{exp(w_k^T x^j)x^j}{1 + \sum_i exp(w_j^T x^j)} \right) - \lambda \sum_{l=1}^{R-1} ||w_l||$$

Problem 1.3

(a) $h = \sum_i a_i g(i)$

(b) $E[g(i)] = P(g(i) = +1) * (+1) + P(g(i) = -1) * (-1) = \frac{1}{2} + -\frac{1}{2} = 0$

(c)

$$E[\hat{a_i}] = E[\sum_j a_j g(j) g(i)]$$

We know that $g(i)$ and $g(j)$ are independent and their expectation is 0. Therefore, in this product, only the terms where $j = i$ are nonzero, leaving just a_i:

$$E[\hat{a_i}] = a_i$$

(d)

$$\begin{aligned}
Var(\hat{a_i}) &= E[\hat{a_i}^2] - (E[\hat{a_i}])^2 \\
&= E[h^2 g(i)^2] - a_i^2 \\
&= E[h^2] - a_i^2 \\
&= E[(\sum_{j=\{1...N\}:j\neq i} a_j g(j))^2] - a_i^2 \\
&= E[\sum_{j\neq i}\sum_{k\neq i} a_j a_k g(j) g(k)] - a_i^2 \\
&= \sum_{j\neq i}\sum_{k\neq i} E[a_j a_k g(j) g(k)] - a_i^{2***} \\
&= \sum_{j=\{1...N\}:j\neq i} a_j^2 - a_i^2 \\
&= \sum_{j=\{1...N\}:j\neq i} a_j^2
\end{aligned}$$

***: Note that all but the diagonals in the double sum term are 0 because $E[g(i)] = 0$.

(e) Using, Chebyshev's inequality, we have:

$$P(|\hat{a_i} - a_i| \geq \epsilon n) \leq \frac{Var(\hat{a_i})}{\epsilon^2 n^2}$$

We need to show $\frac{Var(\hat{a_i})}{\epsilon^2 n^2} \leq \frac{1}{\epsilon^2}$.

$$\frac{Var(\hat{a_i})}{\epsilon^2 n^2} \leq \frac{1}{\epsilon^2}$$

$$\frac{Var(\hat{a_i})}{n^2} \leq 1$$

Prove $n^2 \geq Var(\hat{a_i} = \sum_j a_j^2)$.

We know that a_i can only be incremented or decremented by a count of 1 per element, giving an upper bound as the number of elements (N).

(f) i.

$$\begin{aligned} Var(\hat{a_i}) &= E[\hat{a_i}^2] - (E[\hat{a_i}])^2 \\ &= E[(\frac{1}{k}\sum_{j=1}^{k} h_j g_j(i))^2] - a_i^2 \\ &= \frac{1}{k^2}E[\sum_j \sum_l h_j h_l g_j(i) g_l(i)] - a_i^2 \\ &= \frac{1}{k^2}E[\sum_j h_j^2] - a_i^2 \end{aligned}$$

We know that $\frac{1}{k^2} \leq \frac{1}{k}$ for $k \geq 1$, so the denominator agrees with the bound on the variance. Comparing the numerator, we know $h^2 \leq n^2$ because the accumulator is incremented or decremented by 1 for each sample n. The a_i^2 term is always positive and can only decrease the variance.

Therefore, $Var(\hat{a_i}) \leq \frac{n^2}{k}$.

ii.

$$\begin{aligned} P(|\hat{a}_i - E[\hat{a}_i]| \geq \epsilon n) &\leq \frac{Var(\hat{a_i})}{\epsilon^2 n^2} \\ &\leq \frac{n^2}{\epsilon^2 n^2} \end{aligned}$$

We also have:

$$\begin{aligned} k &\geq \frac{1}{\delta\epsilon^2} \\ \delta\epsilon^2 &\geq \frac{1}{k} \end{aligned}$$

which leads to the result:

$$P(|\hat{a}_i - E[\hat{a}_i]| \geq \epsilon n) \leq \frac{\delta\epsilon^2}{\epsilon^2} \leq \delta$$

Problem 1.4

1. Warm up

(a) Average CTR for training data: 0.033655

(b) 141063 unique tokens in the training set.
109459 unique tokens in the testing set.
79261 unique tokens appear in both datasets.

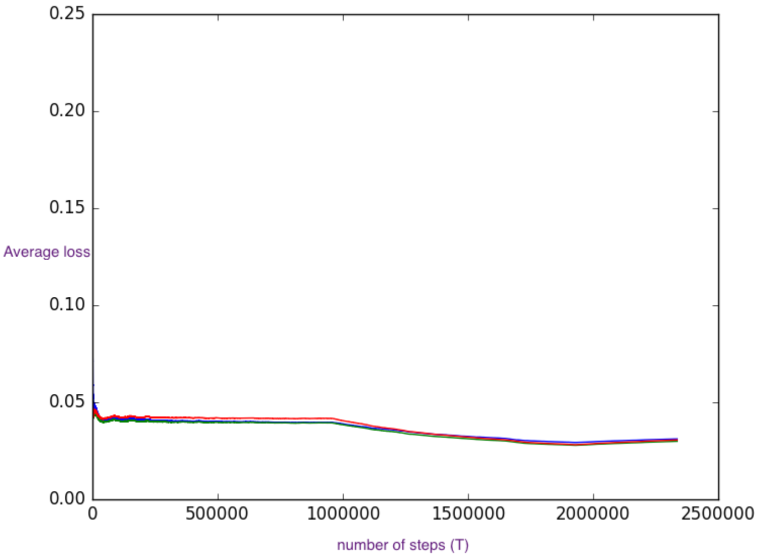
(c) Unique users in:

Age group	Training	Test
0	8826	2021
1	79668	48783
2	162725	113818
3	307482	178416
4	213292	117326
5	146605	79635
6	63834	34908

2. Stochastic Gradient Descent

(a) $w_i^{(t+1)} = w_i^{(t)} + \eta X_i[y_i - P(Y = 1|X^{(t)}, w^{(t)})]$

(b) .

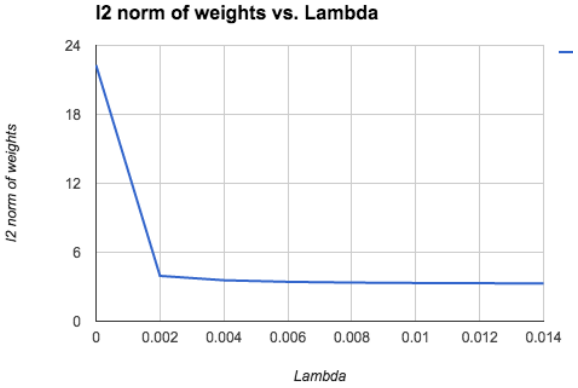


Step size	l2 norm	CTR (test)	RMSE (test)
0.001	3.80187	0.03565	0.17117
0.01	9.05716	0.02616	0.17132
0.05	22.29195	0.03215	0.17340
baseline	–	0.03365	0.17308

(c) For $\eta = 0.01$, we have the following weights:
age: -0.0409480151243
depth: 0.172219387658
gender: 0.104138214772
position: -0.773576136654
Position has the strongest effect, which makes sense; if an ad is located somewhere that happens to harder to see, it will not be clicked on. Depth also has a strong effect because the more ads there are, the more likely one of them will be clicked on (although there is likely and upper and lower bound on this). Gender is slightly biased towards women clicking on more ads, and it is unclear why that may be (perhaps that is a bias in the sample). Age has the least effect, meaning that people of all ages click on ads similarly.

3. Regularization

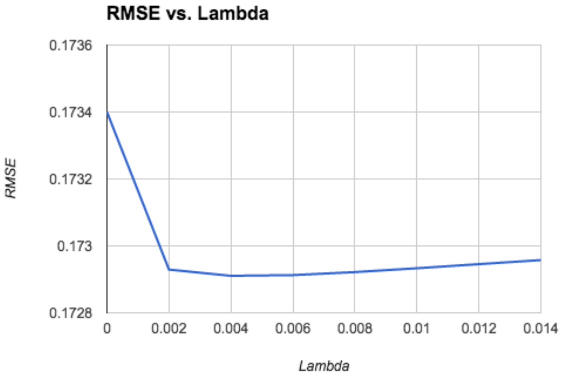
(a) i. .



ii. Yes, there is a consistent trend in the l_2 norm as λ increases – the l_2 norm shrinks and converges to a particular value. This makes sense because the weights converge to the same values with increased regularization.

iii. As we increase $\lambda \to \infty$, the l_2 norm will converge to 0.

(b) .



Lambda	**l2 norm of weights**	**CTR**
0	22.29195	0.03215
0.002	3.94659	0.03914
0.004	3.56506	0.04014
0.006	3.42893	0.040385
0.008	3.36287	0.04045
0.010	3.32655	0.04048
0.012	3.30547	0.0405
0.014	3.29309	0.04052

4. Hashing Kernel

m	**RMSE**
101	0.1725
12277	0.1723
1573549	0.1720