# COMP 652: ASSIGNMENT 1

CARLOS G. OLIVER
260425853

## 1. Regression

1.1. **Q1 (c):Objective for logistic regression with $L_2$ regulariztion.** The logistic regression objective is the cross entropy error function between true outputs $y_i$ and predictions $h(x_i)$ with the additional regularization term proportional to some constant $\lambda$ and the $L_2$ norm of the weight vector $\mathbf{w}$.

$$J_{\mathbf{w}} = -\left( \sum_{i=1}^{m} y_i \log(h(\mathbf{x_i})) + (1 - y_i) \log(1 - h(\mathbf{x_i})) \right) + \frac{\lambda}{2} \mathbf{w}^T \mathbf{w} \tag{1}$$

1.2. **Q1 (d): Regularization.** For all subsequent experiments, I present summary statistics of K-fold cross validation with shuffled input data. We compute a value of $K$ so as to have approximately a training-test split ratio of $80 - 20\%$. This resulted in 4 splits. The bars one each point represent one standard deviation from the mean over the 4 splits.

1.3. **Q1 (e): Gaussian Basis Functions.** Next, we apply a univariate Gaussian basis function $\phi_j(x_i)$ to each value in input vectors $x$ to compute a new feature mapping. Where

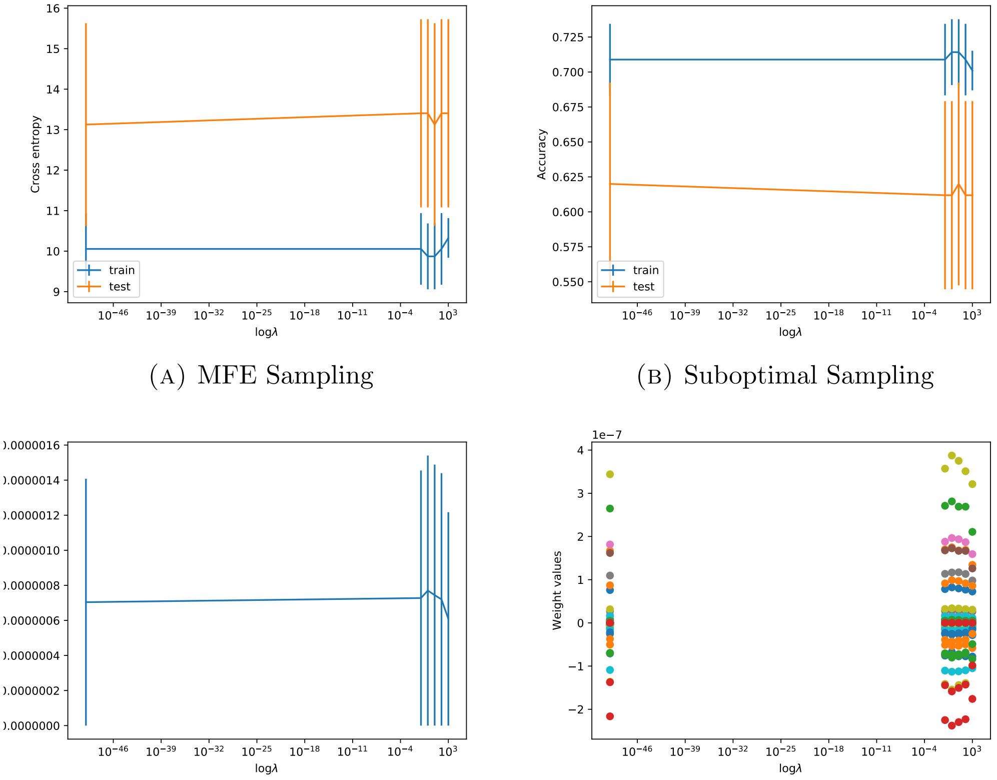
$$\phi(x)_j = exp\left[ \frac{(x_i - \mu_k)^2}{2\sigma^2} \right] \tag{2}$$

We apply this basis with a fixed $\sigma$ chosen from the set $\{0.1, 0.5, 1, 5, 10\}$ at each iteration. For each variable we compute 5 basis functions $(\phi_1(x), ...\phi_5(x))$ with $\mu \in \{-10, -5, 0, 5, 10\}$ resulting in 5 feature vectors for each original feature vector in $X$. This results in a new input matrix $\Phi$.

1.4. **Q1 (f): Effect of $\sigma$ on regression.**

1.5. **Q1 (g): Basis function and regularization.**

*Date*: February 4, 2017.



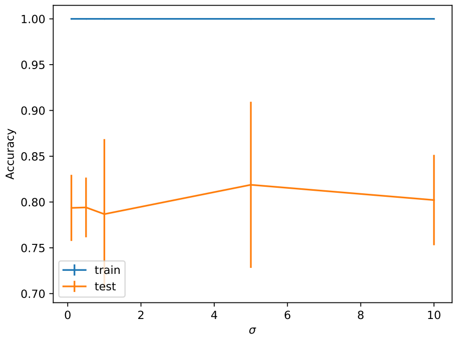
(A) MFE Sampling

(B) Suboptimal Sampling

(C) Suboptimal Sampling

(D) Suboptimal Sampling

FIGURE 1. Frequency of sequence-structure pair in Boltzmann ensemble for MFE and suboptimal sampling in RNAmutants.



1.6. **Q1 (h): Gaussian basis function capturing relationship between inputs.** Instead of using a univariate Gaussian basis function to transform each input variable, we could use multivariate Gaussians to capture the relationship between inputs as covariance. We can use a similar basis as before, but introduce a covariance parameter matrix $\Sigma$ that can be estimated from the data. As we would be increasing the complexity of our model, we are likely to reduce the bias but make the model more sensitive to variations in the data.

1.7. **Q1 (i): Adaptive gaussian center placement.** Previously, we have been externally fixing the centers $\mu_i$ of the Gaussian basis functions. We can instead use the data to adaptively compute the placement of these centers. This would effectively be a clustering task, where each cluster would define the center of a Gaussian basis. With a fixed $\sigma$, the only parameter we need to estimate is the center vector $\mu$ for each center $k$. Our basis functions would then take the form:

$$\phi(\mathbf{x}) = exp\left(\frac{\|\mathbf{x} - \mu\|_2^2}{2\sigma_k^2}\right) \tag{3}$$

```
input  : X, $\lambda$, $\mathbf{y}$, $[\sigma_0, ..., \sigma_j]$
output: $\omega$
$currentTestErr \leftarrow \infty$;
$prevTestErr \leftarrow 0$;
$bestMode \leftarrow Null$;
while $currentTestErr > prevTestErr$ do
    $\mu_1, ..., \mu_j \leftarrow$ KMeans($X$, $k$);
    $X^{'} \leftarrow$ GaussianBasis($X, [\mu_1, ..., \mu_k]$);
    $model, \mathbf{w} \leftarrow$ LogisticRegression($X^{'}, y, \lambda$);
    $currentTestErr \leftarrow$ accuracy($model$, $X$, $y$);
end
```

1.8. **Q1 (j): Convergence of algorithm.**

1.9. **Q2 (a): Dual view of logistic regression.** The hypothesis $h(\mathbf{x})$ in logistic regression takes the form:

$$h(\mathbf{x}) = P(Y = 1|\mathbf{x}, \mathbf{w}) = \frac{1}{1 + e^{\mathbf{w^T x}}} \tag{4}$$

If we wish to obtain the probability distribution of $y$ condition on $\mathbf{x}$ and $\mathbf{w}$ we get:

$$p(y|\mathbf{x}, \mathbf{w}) = \sigma(y\mathbf{w^T x}) = \frac{1}{1 + e^{-y\mathbf{w^T x}}} \tag{5}$$

We can then use this form to find the parameters that maximize the log likelihood of the data given a choice of model. (The following derivation is based on work published by Thomas P. Minka in 2007 [**?**].)

$$l(\mathbf{w}) = \sum_{i=1}^{m} \log \sigma(-y_i \mathbf{w}^{\mathbf{T}} \mathbf{x}) - \frac{\lambda}{2} \mathbf{w}^{\mathbf{T}} \mathbf{w} \tag{6}$$

In order to derive a dual expression of this objective, we aim to find a tight linear upper bound parametrized by a new set of parameters $\alpha_i$. Such a function will let us reverse the order of the optimization while maintaining the same optima. In our case, if we wish to maximize $l(\mathbf{w})$ we can instead minimize some function $l(\mathbf{w}, \alpha)$ that is an upper bound to the first. We can then find the minimal value of $l(\mathbf{w}, \alpha)$ that maximizes $l(\mathbf{w})$.

We use the function:

$$H(\alpha) = -\alpha \log \alpha - (1 - \alpha) \log(1 - \alpha), \quad \alpha \in [0, 1] \tag{7}$$

Plotting $H(x)$ we see that the function is quadratic.

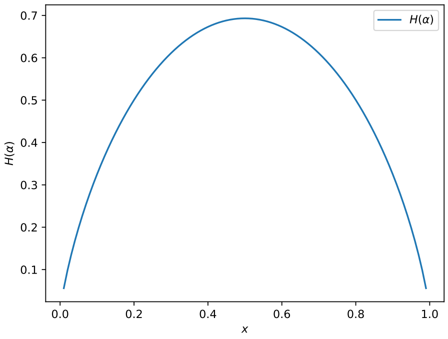


FIGURE 2. Parameter function

Using $H(x)$ we can bound $\log(\sigma(x))$ above as follows:

$$\log \sigma(x) \leq \alpha x - H(x) \tag{8}$$

Plotting these functions with $\alpha = \{0, 0.5, 1\}$ we see that indeed, $\log \sigma(x)$ is tightly bounded above.

We can then apply these constraints to the original log likelihood function optimization problem:

$$l(\mathbf{w}) = \sum_{i}^{m} \log \sigma(y_i \mathbf{w}^{\mathbf{T}} \mathbf{x}) - \frac{\lambda}{2} \mathbf{w}^{\mathbf{T}} \mathbf{w} \leq l(\mathbf{w}, \alpha) \tag{9}$$

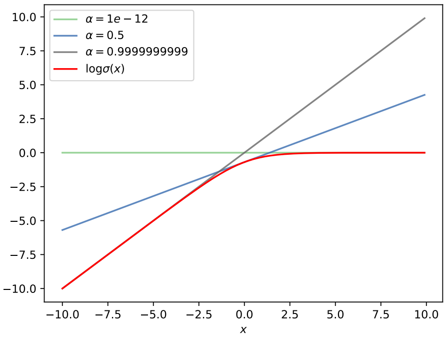


FIGURE 3. Upper bounds on log likelihood function

$$\text{Where} \quad l(\mathbf{w}, \alpha) = \sum_{i}^{m} \alpha_i y_i \mathbf{w^T x_i} - H(\alpha_i) - \frac{\lambda}{2}\mathbf{w^T w} \tag{10}$$

We can now write the problem as

$$\min_{\alpha} \max_{\mathbf{w}} l(\mathbf{w}, \alpha) \tag{11}$$

Taking the derivative of $l(\mathbf{w}, \alpha)$ with respect to $\mathbf{w}$ and setting to zero we get an expression for $\mathbf{w}$

$$\mathbf{w} = -\lambda \sum_{i}^{m} \alpha_i y_i \mathbf{x_i} \tag{12}$$

Which we can plug into the original $l(\mathbf{w}, \alpha$ and now optimize with respect to $\alpha$ to obtain

$$J(\alpha) = \frac{1}{2\lambda} \sum_{i,j}^{m,n} \alpha_i \alpha_j y_i y_j \mathbf{x_j^T x_i} - \sum_{i}^{m} H(\alpha_i) \tag{13}$$

## REFERENCES

[1] Thomas P Minka. A comparison of numerical optimizers for logistic regression. *Unpublished draft*, 2003.