Questions for Flipped Classroom Session of COMS 4705  
Week 11, Fall 2014. (Michael Collins)

**Question 1**  
In this question we consider the problem of mapping a sentence to an underlying sequence of tags, using a log-linear tagger. The input to the tagger is a sequence of words $x_1 x_2 \ldots x_N$, where each $x_i$ is an English word. The output from the tagger is a sequence of tags $y_1 y_2 \ldots y_N$. Each tag $y_i$ can take any one of four possible states, $A$, $B$, $C$ or $D$. Looking at the data, we notice that tag sequences follow the following rules:

- The tag $y_1$ is always equal to either $A$ or $B$.
- For all tag bigrams $y_j, y_{j+1}$, we either have $y_j \in \{A, B\}$, and $y_{j+1} \in \{C, D\}$; or we have $y_j \in \{C, D\}$, and $y_{j+1} \in \{A, B\}$. (That is, we never see the tag bigrams $AA$, $AB$, $BA$, $BB$, $CC$, $CD$, $DC$ or $DD$.)
- If the $j$’th word $x_j$ in the sequence has an odd number of letters, its tag $y_j$ is always equal to either $A$ or $C$. If the $j$’th word has an even number of letters, its tag is always equal to either $B$ or $D$.

We will use a log-linear bigram tagger to map sentences to tag sequences. The model takes the form

$$ p(y_1 \ldots y_N | x_1 \ldots x_N) = \prod_{j=1}^N p(y_j | y_{j-1}, x_1 \ldots x_N, j) $$

where

$$ p(y_j | y_{j-1}, x_1 \ldots x_N, j) = \frac{\exp\{v \cdot f(y_{j-1}, x_1 \ldots x_N, j, y_j)\}}{\sum_{y \in \{A, B, C, D\}} \exp\{v \cdot f(y_{j-1}, x_1 \ldots x_N, j, y)\}} $$

and $f(y_{j-1}, x_1 \ldots x_N, j, y) \in \mathbb{R}^d$ is a feature vector, and $v \in \mathbb{R}^d$ is a parameter vector, where $d$ is the number of parameters.

**Question:** Give a feature-vector definition $f(y_{j-1}, x_1 \ldots x_N, j, y) \in \mathbb{R}^d$ that allows the model to perfectly model the constraints given above.
**Question 2** Say we are running the perceptron algorithm. We have reached example \( x_i \) and the set \( \{ f(x_i, y) : y \in \text{GEN}(x_i) \} \) consists of the following vectors:

(a) \( \langle 1, 0, 1, 1 \rangle \)
(b) \( \langle 1, 1, 1, 0 \rangle \)
(c) \( \langle 0, 0, 1, 1 \rangle \)

Assume also that \( f(x_i, y_i) = \langle 0, 0, 1, 1 \rangle \).

**Question:** Give a setting for the parameter vector \( v \) that ensures that the output of the global linear model on \( x_i \) is \( y_i \).

**Question:** Now assume that \( v = \langle 1, 1, -1, -1 \rangle \) immediately before this example is considered by the algorithm. What will the value of \( v \) be at the end of this iteration?

**Question:** Now assume that \( v = \langle 1, 1, -1, -1 \rangle \), and we run the perceptron algorithm repeatedly on the example above. What parameter values does the algorithm converge to? Assume that when computing

\[
\arg \max_{y \in \text{GEN}(x_i)} v \cdot f(x_i, y)
\]

any ties in the score \( v \cdot f(x_i, y) \) are broken in the order (a) > (b) > (c).
Figure 1: A modified version of the perceptron algorithm.

**Question 3** Figure 1 shows a modified version of the perceptron algorithm. Show that under the same definitions for $\delta$ and $R$ for the regular perceptron, the algorithm makes at most

$$
\frac{R^2}{\delta^2}
$$

updates to $v$ before convergence. (See over the page for the proof of convergence for the regular perceptron algorithm.)
Appendix to Question 3: Proof of Convergence for the Perceptron Algorithm

• **Definition:** \( \text{GEN}(x_i) = \text{GEN}(x_i) - \{y_i\} \)

• **Definition:** The training set is **separable with margin** \( \delta \), if there is a vector \( u \in \mathbb{R}^d \) with \( ||u||_2 = 1 \) such that

\[
\forall i, \forall z \in \text{GEN}(x_i) \quad u \cdot f(x_i, y_i) - u \cdot f(x_i, z) \geq \delta
\]

**Theorem:** For any training sequence \( (x_i, y_i) \) which is separable with margin \( \delta \), then for the perceptron algorithm

\[
N \leq \frac{R^2}{\delta^2}
\]

where \( N \) is the number of updates to \( v \), \( R \) is a constant such that \( \forall i, \forall z \in \text{GEN}(x_i) \quad ||f(x_i, y_i) - f(x_i, z)||_2 \leq R \)

**Proof:** Direct modification of the proof for the classification case.

Let \( v^k \) be the weights before the \( k \)'th mistake. \( v^1 = 0 \)

If the \( k \)'th mistake is made at \( i \)'th example, and \( z_i = \arg\max_{y \in \text{GEN}(x_i)} f(x_i, y) \cdot v^k \), then

\[
\begin{align*}
v^{k+1} &= v^k + f(x_i, y_i) - f(x_i, z_i) \\
\Rightarrow u \cdot v^{k+1} &= u \cdot v^k + u \cdot f(x_i, y_i) - u \cdot f(x_i, z_i) \\
&\geq u \cdot v^k + \delta \\
&\geq k \delta
\end{align*}
\]

Also,

\[
\begin{align*}
||v^{k+1}||_2^2 &= ||v^k||_2^2 + ||f(x_i, y_i) - f(x_i, z_i)||_2^2 + 2v^k \cdot (f(x_i, y_i) - f(x_i, z_i)) \\
&\leq ||v^k||_2^2 + R^2 \\
\Rightarrow ||v^{k+1}||_2^2 &\leq kR^2 \\
\Rightarrow k^2 \delta^2 &\leq ||v^{k+1}||_2^2 \leq kR^2 \\
\Rightarrow k &\leq \frac{R^2}{\delta^2}
\end{align*}
\]