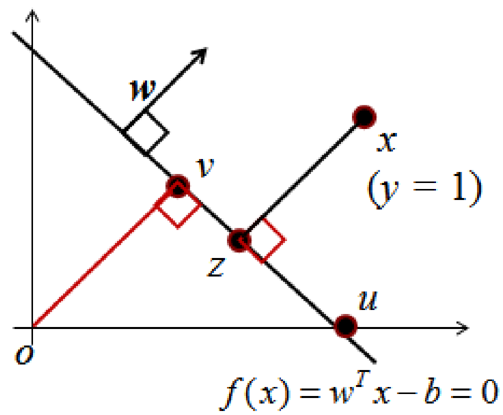
11-741/11-641/11-441 – Machine Learning for Text Mining **Due: Jan 21, 2016, 11:59 PM**
Instructor: Yiming Yang

Homework 1

TA in charge: Pengfei Wang

This assignment consists of a set of problems, which will give you the experience with relevant concepts in later lectures of this course.

1. **Hyperplane properties** (useful concept in defining Support Vector Machines) **[25pt]**
   a) **[10pt]** Prove that the shortest distance from the origin to hyperplane h is $\frac{|b|}{\|w\|}$ where
   $$h = \left\{ x : w^T x - b = 0 \right\},\ w \in R^m,\ b \in R,\ x = (x_1, x_2, \cdots, x_m).$$
   [Hint: use the cosine of the angle between normal vector w and the horizontal axis in the figure.]



   b) **[15pt]** Prove that the perpendicular distance from point x to h is $\frac{yf(x)}{\|w\|} = \frac{y(w^T x - b)}{\|w\|}$ where $y \in \{-1,1\}$ is the indicator of which side of h vector x is located.

   References:
   - C.M. Bishop, Pattern Recognition and Machine Learning, page 181-182
   - Youtube, e.g., http://www.youtube.com/watch?v=AUzQg79gKJQ

2. **Eigenvalues and eigenvectors** (useful for proving the convergence of PageRank, etc.) **[20pt]**
   a) **[10pt]** Find the eigenvalues and eigenvectors of $A = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$; show your calculation steps.
   b) **[10pt]** Show that the eigenvalues of A^k are $\lambda_1^k, \lambda_2^k, \cdots, \lambda_n^k$, the kth powers of the eigenvalues of matrix A, and that each eigenvector of A is still an eigenvector of A^k.

   References:

- Wikipedia: http://en.wikipedia.org/wiki/Eigenvalues_and_eigenvectors
- Gilbert Strang. Linear Algebra and its Applications, Ch 5. HBJ Publishers.

3. **Maximum likelihood estimate** (related to probabilistic models) **[25pt]**

   a) **[10pt]** Prove that in a binomial process of coin tossing, if the count of head-up is *k* out of *n*, then the maximum likelihood estimate (MLE) of the true head-up probability is $\hat{p} = \frac{k}{n}$.

   b) **[15pt]** Prove that in a multinomial process of sampling *n* words from a vocabulary of size *m*, the relative frequency of each word is the MLE of the true probability of that word:

   $$\hat{p}_j = \frac{n_j}{n}, \forall j = 1,2,\cdots m,$$

   where n_j is the count of a specific word, and $n_1 + \cdots + n_m = n$.

4. **Calculus (**related to parameter optimization in logistic regression for classification) **[30pt + extra 15pt]**

   a) **[10pt]** Show how to calculate the 1st derivative of sigmoid function $u = \frac{1}{1+e^{-x}}$ with respect to x, i.e., prove that $\frac{du}{dx} = u(1-u)$.

   b) **[10pt]** A multivariate function is defined as $l = y \ln u + (1-y)\ln(1-u)$ where $u = \frac{1}{1+e^{-z}}$, $z = w^T x = w_0 + w_1 x_1 + \cdots + w_m x$ and $w = (w_0, w_1, \cdots, w_m)$. Assuming x and y are given, show the calculation of the gradient of *l* w.r.t. $w_0, w_1, \cdots, w_m$:

   $$\nabla l = \left(\frac{\partial l}{\partial w_0}, \quad \frac{\partial l}{\partial w_1}, \quad \cdots, \quad \frac{\partial l}{\partial w_m} \right)^T$$

   c) **[10pt]** Show the calculation of the pairwise 2nd order derivative $H_{jj'} \equiv \frac{\partial^2 l}{\partial w_j \partial w_{j'}}$.

   d) **[Extra 15pt]** Show that the log-likelihood function in logistic regression is concave.

Turn In: Written Report

Submit your report in **PDF** format and name it as " HW1-*YourAndrewID*.pdf ". A typeset solution is preferred, but you may alternatively scan a handwritten solution. Please also include your **name and Andrew ID** at the **top of the first page** of your report. Your report must contain the solutions for all of the problems, with **each subproblem clearly labeled as an independent section**.

Restrictions

1. You must write all of the solutions yourself.
2. You must show your work for full credit.
3. If you submit a scanned handwritten solution, make sure that your handwriting is clear and legible. **If the TA cannot read your handwriting, the answer will be assumed to be incorrect.**
