# Midterm Exam Review
# Part 1

Xi Chen

# TRUE/FALSE QUESTIONS

- For a continuous random variable x and its probability distribution function p(x), it holds that $0<=p(x)<=1$ for all x

- The error of a hypothesis measured over its training set provides a pessimistically biased estimate of the true error of the hypothesis.

- Given m data points, the training error converges to the true error as $m \rightarrow \infty$.

- Maximizing the likelihood of logistic regression model yields multiple local optimums.

- No classifier can do better than a naive Bayes classifier if the distribution of the data is known.

- The correspondence between logistic regression and Gaussian Na�e Bayes (with identity class covariances) means that there is a one-to-one correspondence between the parameters of the two classifiers

# Regression and bias/variance tradeoff

Suppose you have regression by a polynomial of degree 3. Characterize the bias-variance of the estimates of the following models on the data:

| | Bias | Variance |
|---|---|---|
| Linear Regression | Low / High | Low / High |
| Polynomial Regression with degree 3 | Low / High | Low / High |
| Polynomial Regression with degree 10 | Low / High | Low / High |

# Decision Tree Question (Spring 2006 Midterm Problem 2)

The following data set will be used to learn a decision tree for predicting whether students are lazy (L) or diligent (D) based on their weight (Normal or Underweight), their eye color (Amber or Violet) and the number of eyes they have (2 or 3 or 4).

| Weight | Eye Color | Num. Eyes | Output |
|---|---|---|---|
| N | A | 2 | L |
| N | V | 2 | L |
| N | V | 2 | L |
| U | V | 3 | L |
| U | V | 3 | L |
| U | A | 4 | D |
| N | A | 4 | D |
| N | V | 4 | D |
| U | A | 3 | D |
| U | A | 3 | D |

The following numbers may be helpful as you answer this problem without using a calculator: $\log_2 0.1 = -3.32, \log_2 0.2 = -2.32, \log_2 0.3 = -1.73, \log_2 0.4 = -1.32, \log_2 0.5 = -1$.

1. (3 pts) What is the conditional entropy $H(EyeColor|Weight = N)$?
2. (3 pts) What attribute would the ID3 algorithm choose to use for the root of the tree (no pruning)?
3. (4 pts) Draw the full decision tree learned for this data (no pruning).
4. (2 pts) What is the training set error of this unpruned tree?

# Logistic Regression Problem (Fall 2010 Midterm Problem 3)

We consider here a discriminative approach for solving the classification problem illustrated in Figure 1.

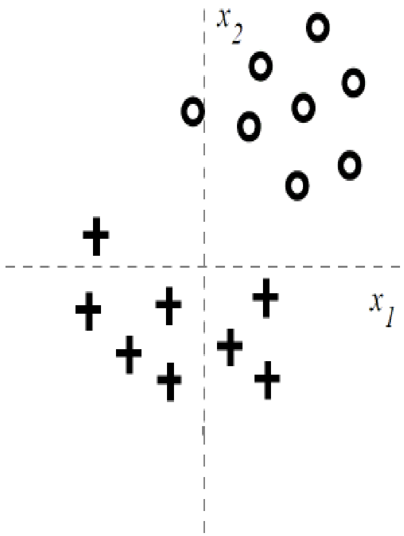


Figure 1: The 2-dimensional labeled training set, where '+' corresponds to class $y$=1 and 'O' corresponds to class $y = 0$.

1. We attempt to solve the binary classification task depicted in Figure 1 with the simple linear logistic regression model

$$P(y = 1|\vec{x}, \vec{w}) = g(w_0 + w_1x_1 + w_2x_2) = \frac{1}{1 + exp(-w_0 - w_1x_1 - w_2x_2)}.$$

Notice that the training data can be separated with *zero* training error with a linear separator.

Consider training *regularized* linear logistic regression models where we try to maximize

$$\sum_{i=1}^{n} \log\left(P(y_i|x_i, w_0, w_1, w_2)\right) - Cw_j^2$$

(a) Regularizing w_2, training error (increase, remain the same, decrease)?
(b) Regularizing w_1, training error?
(c) Regularizing w_0, training error?

2. If we change the form of regularization to L1-norm (absolute value) and regularize $w_1$ and $w_2$ only (but not $w_0$), we get the following penalized log-likelihood

$$\sum_{i=1}^{n} \log P(y_i|x_i, w_0, w_1, w_2) - C(|w_1| + |w_2|).$$

(a) [3 pts] As we increase the regularization parameter $C$ which of the following scenarios do you expect to observe? (Choose only one) Briefly explain your choice:

( ) First $w_1$ will become 0, then $w_2$.
( ) First $w_2$ will become 0, then $w_1$.
( ) $w_1$ and $w_2$ will become zero simultaneously.
( ) None of the weights will become exactly zero, only smaller as $C$ increases.

(b) [3 pts] For very large $C$, with the same L1-norm regularization for $w_1$ and $w_2$ as above, which value(s) do you expect $w_0$ to take? Explain briefly. (Note that the number of points from each class is the same.) (You can give a range of values for $w_0$ if you deem necessary).