6.867 Machine learning

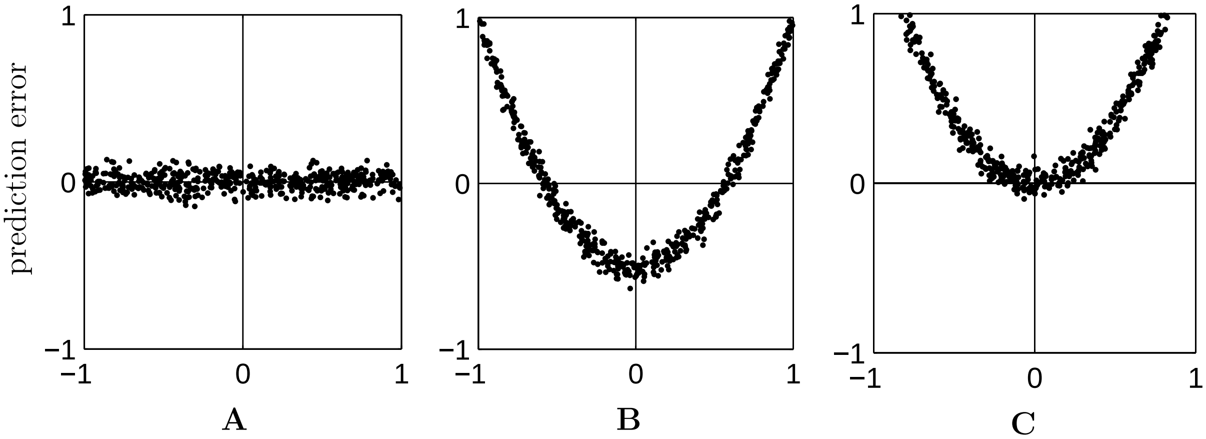
Mid-term exam

October 13, 2004

(2 points) Your name and MIT ID:

T. Assistant, 968672004

Problem 1



1. **(6 points)** Each plot above claims to represent prediction errors as a function of x for a trained regression model based on some dataset. Some of these plots could potentially be prediction errors for linear or quadratic regression models, while others couldn't. The regression models are trained with the least squares estimation criterion. Please indicate compatible models and plots.

	A	**B**	**C**
linear regression	(x)	(x)	( )
quadratic regression	(x)	( )	( )

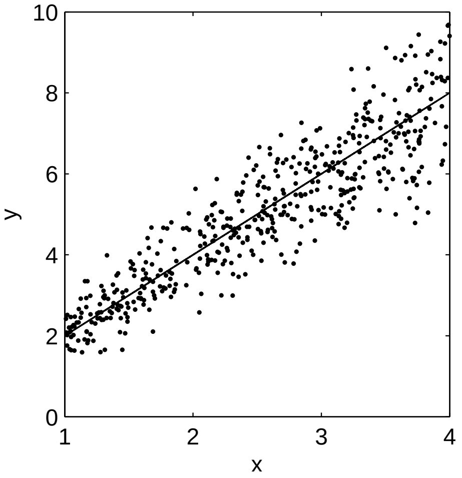
Problem 2

Here we explore a regression model where the noise variance is a function of the input (variance increases as a function of input). Specifically

$$y = wx + \epsilon$$

where the noise ϵ is normally distributed with mean 0 and standard deviation σx. The value of σ is assumed known and the input x is restricted to the interval $[1, 4]$. We can write the model more compactly as $y \sim N(wx, \sigma^2 x^2)$.

If we let x vary within $[1, 4]$ and sample outputs y from this model with some w, the regression plot might look like



1. **(2 points)** How is the ratio y/x distributed for a fixed (constant) x?

   Since $y \sim N(wx, \sigma^2 x^2)$, for any constant x, y/x is also Gaussian with mean $wx/x = w$ and variance $\sigma^2 x^2 / x^2 = \sigma^2$. So, $y/x \sim N(w, \sigma^2)$.

2. Suppose we now have n training points and targets $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where each x_i is chosen at random from $[1, 4]$ and the corresponding y_i is subsequently sampled from $y_i \sim N(w^* x_i, \sigma^2 x_i^2)$ with some true underlying parameter value w^*; the value of σ^2 is the same as in our model.

(a) **(3 points)** What is the maximum-likelihood estimate of w as a function of the training data?

We know that $y/x \sim N(w, \sigma^2)$. We can therefore estimate w by interpreting y_i/x_i as observations. The maximum likelihood estimate of w is simply the mean

$$\hat{w}_n = \frac{1}{n} y_i/x_i$$

(b) **(3 points)** What is the variance of this estimator due to the noise in the target outputs as a function of n and σ^2 for fixed inputs $x_1, \ldots, x_n$? For later utility (if you omit this answer) you can denote the answer as $V(n, \sigma^2)$.

The variance of the estimator of a mean of a gaussian is σ^2/n.

Some potentially useful relations: if $z \sim N(\mu, \sigma^2)$, then $az \sim N(a\mu, \sigma^2 a^2)$ for a fixed a. If $z_1 \sim N(\mu_1, \sigma_1^2)$ and $z_2 \sim N(\mu_2, \sigma_2^2)$ and they are independent, then $\mathrm{Var}(z_1 + z_2) = \sigma_1^2 + \sigma_2^2$.

3. In sequential active learning we are free to choose the next training input x_{n+1}, here within $[1, 4]$, for which we will then receive the corresponding noisy target y_{n+1}, sampled from the underlying model. Suppose we already have $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ and are trying to figure out which x_{n+1} to select. The goal is to choose x_{n+1} so as to help minimize the variance of the predictions $f(x; \hat{w}_n) = \hat{w}_n x$, where $\hat{w}_n$ is the maximum likelihood estimate of the parameter w based on the first n training examples.

(a) **(2 points)** What is the variance of $f(x; \hat{w}_n)$ due to the noise in the training outputs as a function of x, n, and σ^2 given fixed (already chosen) inputs $x_1, \ldots, x_n$?

$f(x; \hat{w}_n)$ is xw, and we know that the variance of w is σ^2/n (from the previous part). Thus the variance of $f(x; \hat{w}_n)$ is $x^2\sigma^2/n$.

(b) **(2 points)** Which x_{n+1} would we choose (within $[1, 4]$) if we were to next select x with the maximum variance of $f(x; \hat{w}_n)$?

The variance is maximized when x^2 is maximum, that is $x = 4$.

(c) **(T/F – 2 points)** Since the variance of $f(x; \hat{w}_n)$ only depends on x, n, and σ^2, we could equally well select the next point at random from $[1, 4]$ and obtain the same reduction in the maximum variance.

T

The variance at $x = 4$ is $16\sigma^2/n$. This does not depend on the actual choice of the queried points, but only on the number of points queried.

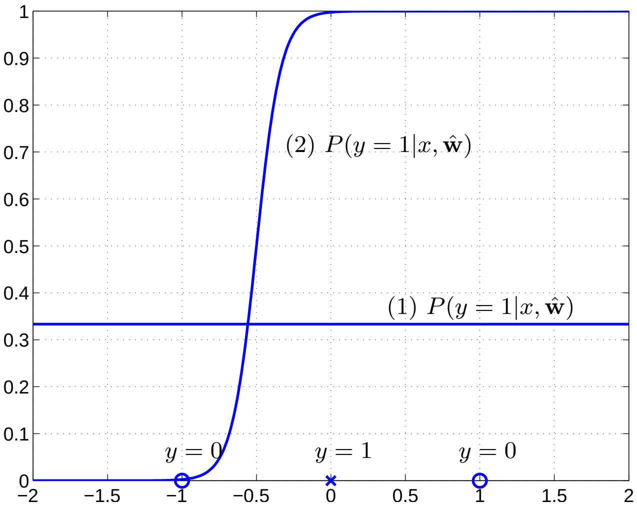


Figure 1: Two possible logistic regression solutions for the three labeled points.

Problem 3

Consider a simple one dimensional logistic regression model

$$P(y = 1|x, \mathbf{w}) = g(w_0 + w_1 x)$$

where $g(z) = (1 + \exp(-z))^{-1}$ is the logistic function.

1. Figure 1 shows two possible conditional distributions $P(y = 1|x, \mathbf{w})$, viewed as a function of x, that we can get by changing the parameters $\mathbf{w}$.

   (a) **(2 points)** Please indicate the number of classification errors for each conditional given the labeled examples in the same figure

   Conditional (1) makes (*1*) classification errors
   Conditional (2) makes (*1*) classification errors

   (b) **(3 points)** One of the conditionals in Figure 3 corresponds to the maximum likelihood setting of the parameters $\hat{\mathbf{w}}$ based on the labeled data in the figure. Which one is the ML solution (1 or 2)?

   $\boxed{1}$

   The likelihood under model (2) is 0*, because model (2) assigns 0 probability to the sample at* 1*. The likelihood under model (1) is* $2/3 \cdot 1/3 \cdot 2/3$*.*

   (c) **(2 points)** Would adding a regularization penalty $|w_1|^2/2$ to the log-likelihood estimation criterion affect your choice of solution (Y/N)?

   $\boxed{\text{N}}$

   At maximum likelihood $\hat{w}_1 = 0$ *(cf. Conditional (1)), thus* $|\hat{w}_1|^2/2$ *is minimal even without a regularization penalty.*

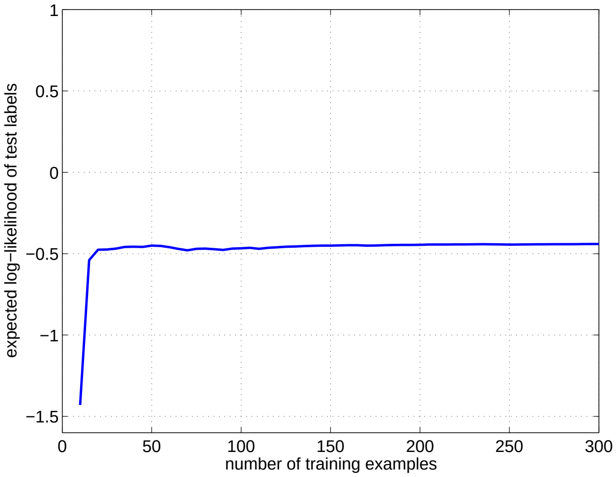


Figure 2: The expected log-likelihood of test labels as a function of the number of training examples.

2. **(4 points)** We can estimate the logistic regression parameters more accurately with more training data. Figure 2 shows the expected log-likelihood of test labels for a simple logistic regression model as a function of the number of training examples and labels. *Mark in the figure* the structural error (SE) and approximation error (AE), where "error" is measured in terms of log-likelihood.

   SE is the distance from the horizontal part of the graph to $y = 0$. AE is everything below the horizontal part of the graph.

3. **(T/F – 2 points)** In general for small training sets, we are likely to reduce the approximation error by adding a regularization penalty $|w_1|^2/2$ to the log-likelihood criterion. T

   Regularization prevents over-fitting by constraining the input to the logistic function to be a smooth function of input. Such functions can be estimated with fewer samples. Put another way, you are reducing the variance of the estimator in favor of introducing some bias. Variance dominates when we have only few training samples.

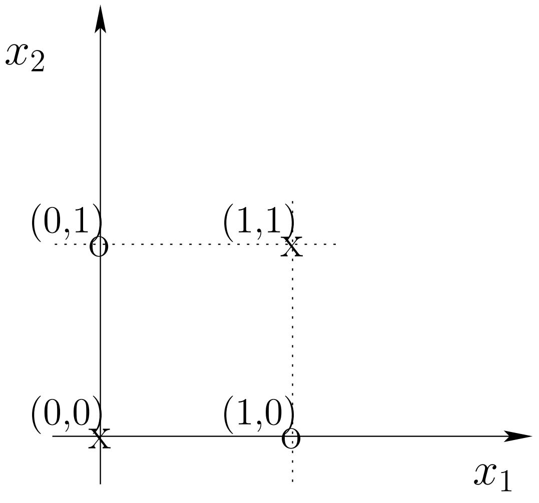


Figure 3: Equally likely input configurations in the training set

Problem 4

Here we will look at methods for selecting input features for a logistic regression model

$$P(y = 1|\mathbf{x}, \mathbf{w}) = g(w_0 + w_1 x_1 + w_2 x_2)$$

The available training examples are very simple, involving only binary valued inputs:

Number of copies	x_1	x_2	y
10	1	1	1
10	0	1	0
10	1	0	0
10	0	0	1

So, for example, there are 10 copies of $\mathbf{x} = [1, 1]^T$ in the training set, all labeled $y = 1$. The correct label is actually a deterministic function of the two features: $y = 1$ if $x_1 = x_2$ and zero otherwise.

We define greedy selection in this context as follows: we start with no features (train only with w_0) and successively try to add new features provided that each addition strictly improves the training log-likelihood. We use no other stopping criterion.

1. **(2 points)** Could greedy selection add either x_1 or x_2 in this case? Answer Y or N.

   N

   We have equally many $y = 1$ and $y = 0$ examples under $x_1 = 1$, and the same is true for $x_1 = 0$. Thus we cannot bias the probabilities in any way based on the presence of x_1 alone. The same is true for x_2.

2. **(2 points)** What is the classification error of the training examples that we could achieve by including both x_1 and x_2 in the logistic regression model? $\boxed{0.25}$

3. **(3 points)** Suppose we define another possible feature to include, a function of x_1 and x_2. Which of the following features, if any, would permit us to correctly classify all the training examples when used in combination with x_1 and x_2 in the logistic regression model:

   ( ) $x_1 - x_2$
   (x) x_1x_2
   ( ) x_2^2

   $x_1 - x_2$ is a linear combination of existing features, thus adding it does not change the model in any way. The same is true for x_2^2, because $x_2^2 = x_2$ for $x_2 \in \{0, 1\}$.
   Now suppose we model the conditional by $P(y = 1|\mathbf{x}, \mathbf{w}) = g(w_0 + w_1x_1 + w_2x_2 + w_3x_1x_2)$. Let w_1 and w_2 be such that when $w_3 = 0$ only $(1, 1)$ is misclassified. Since $x_1x_2 = 0$ except at $(1, 1)$, w_3 does not affect the classification of the correctly classified training points. We can than choose w_3 such that $(1, 1)$ is also correctly classified.

4. **(2 points)** Could the greedy selection method choose this feature as the first feature to add when the available features are x_1, x_2 and your choice of the new feature? Answer Y or N. $\boxed{\text{Y}}$

Problem 5

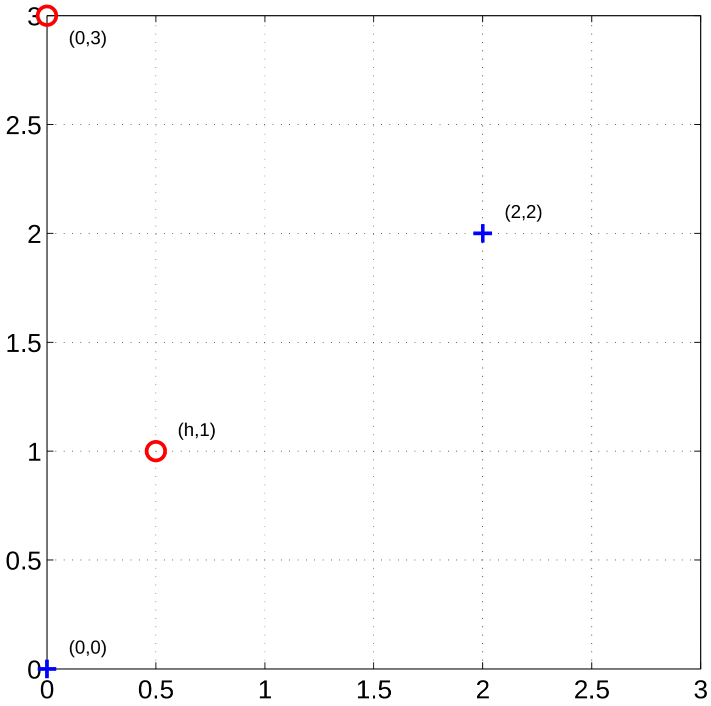


Figure 4: Labeled training examples

Suppose we only have four training examples in two dimensions (see Figure 4):

positive examples at $\mathbf{x}_1 = [0,0]^T, \mathbf{x}_2 = [2,2]^T$ and
negative examples at $\mathbf{x}_3 = [h,1]^T, \mathbf{x}_4 = [0,3]^T$.

where we treat $0 \leq h \leq 3$ as a parameter.

1. **(2 points)** How large can $h \geq 0$ be so that the training points are still linearly separable? $\boxed{h \leq 1}$

2. **(2 points)** Does the orientation of the maximum margin decision boundary change as a function of h when the points are separable? Answer Y or N. $\boxed{\text{N}}$

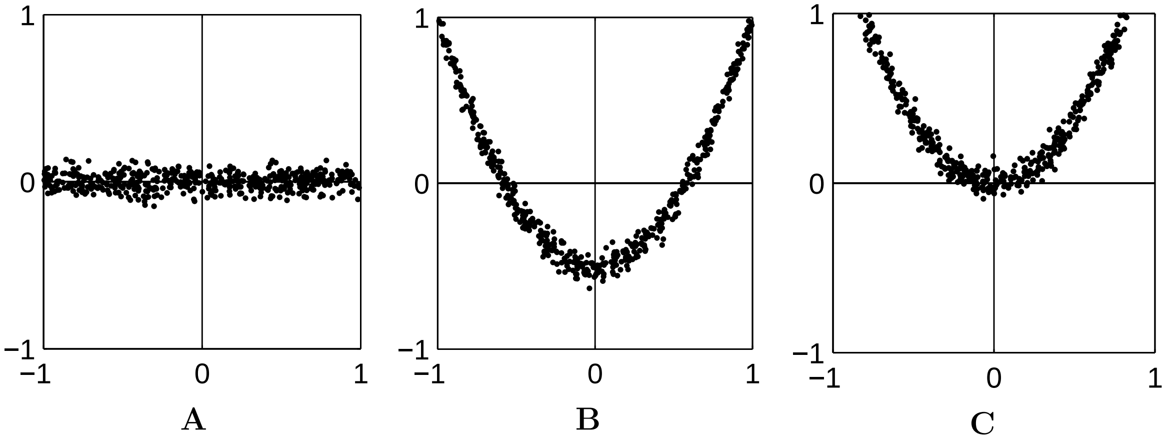
3. **(4 points)**What is the margin achieved by the maximum margin boundary as a function of h?

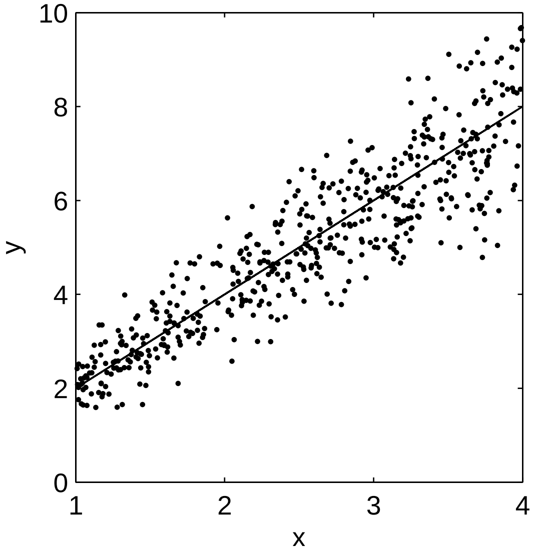
The margin is $(1-h)/\sqrt{2}$*, for* $0 \leq h \leq 1$*, and* 0 *for* $h > 1$*.*
One way to calculate the margin is as follows: since the orientation of the boundary does not change as a function h*, the margin is a linear function of* h*. The margin is zero at* $h = 1$ *and* $1/\sqrt{2}$ *at* $h = 0$*.*

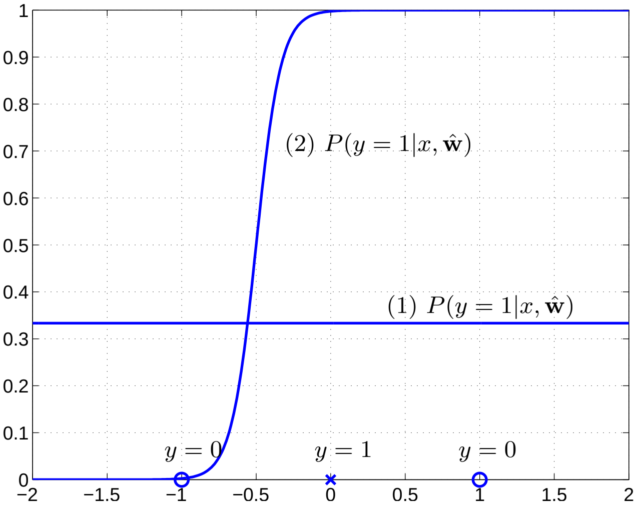
4. **(3 points)** Assume that $h = 1/2$ (as in the figure) and that we can only observe the x_2-component of the input vectors. Without the other component, the labeled training points reduce to $(0, y = 1)$, $(2, y = 1)$, $(1, y = -1)$, and $(3, y = -1)$. What is the lowest order p of polynomial kernel that would allow us to correctly classify these points?

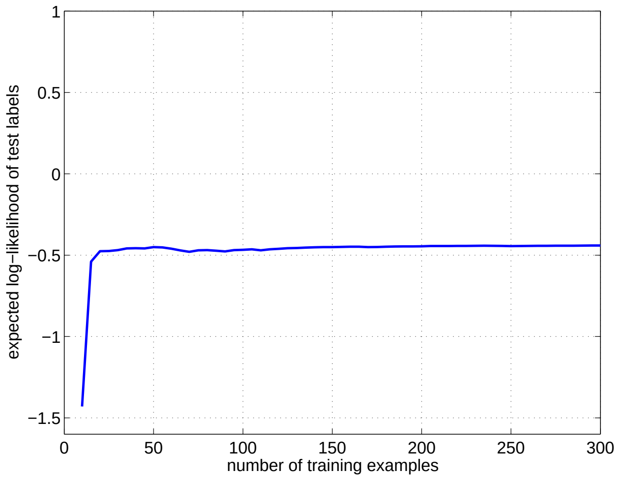
3

Additional set of figures

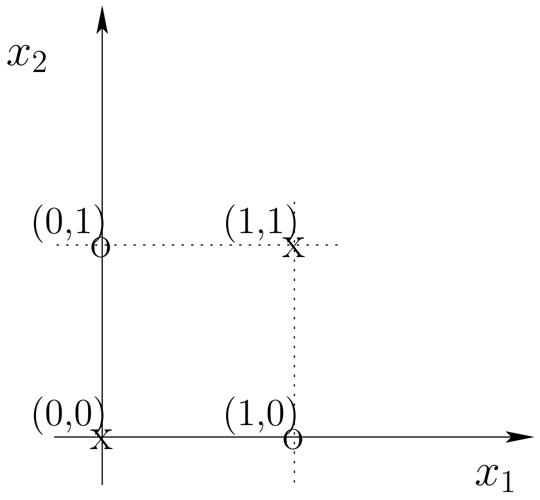




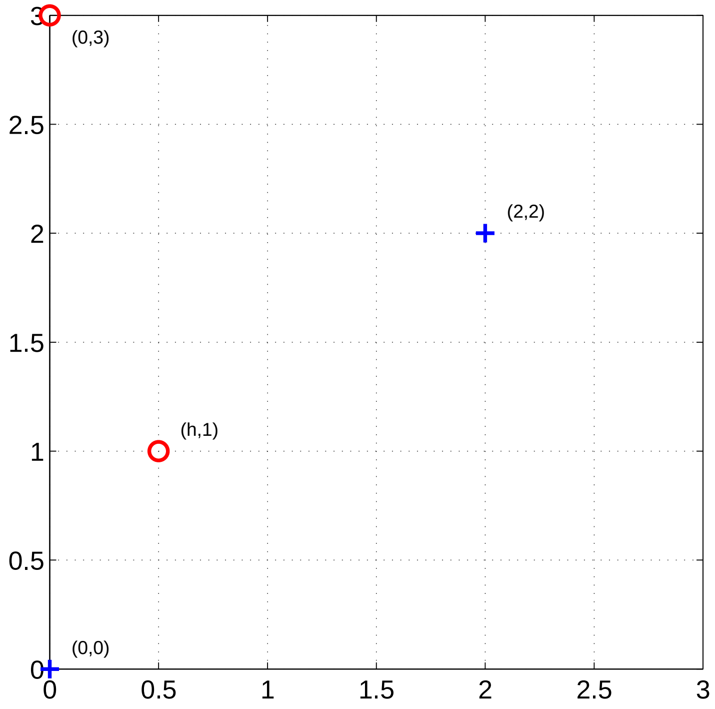




expected log−likelihood of test labels
1
0.5
0
−0.5
−1
−1.5
0
50
100
150
200
250
300
number of training examples



x_2
(0,1)
(1,1)
o
x
(0,0)
(1,0)
x
o
x_1



3
(0,3)
2.5
(2,2)
2
1.5
(h,1)
1
0.5
(0,0)
0
0
0.5
1
1.5
2
2.5
3