(f) In one sentence, characterize the differences between *maximum likelihood* and *maximum a posteriori* approaches.

maximum likelihood finds parameters to maximizes the likelihood function, while MAP maximizes the posterior probability

(g) In one sentence, characterize the differences between classification and regression.

classification maps inputs to discrete outputs
regression maps inputs to continuous outputs

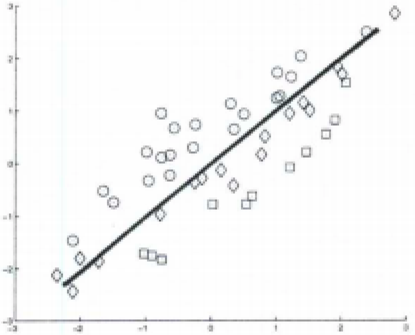
(h) Give one similarity and one difference between feature selection and PCA.

similarity: reduce the dimension of data
difference: feature selection finds a subset of features, while PCA

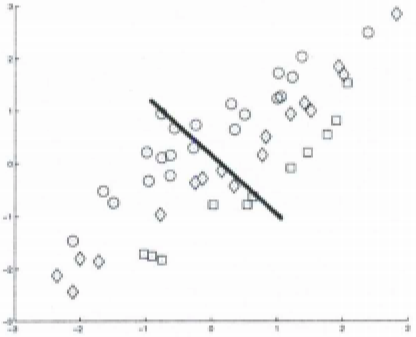
7 Dimensionality Reduction (8pts)

In this problem four linear dimensionality reduction methods will be discussed. They are principal component analysis (PCA), linear discriminant analysis (LDA), canonical correlation analysis (CCA), non-negative matrix factorization (NMF).

1. (3pts) LDA reduces the dimensionality given labels by *maximizing the overall interclass variance relative to intraclass variance.* Plot the directions of the *first* PCA and LDA components in the following figures respectively.

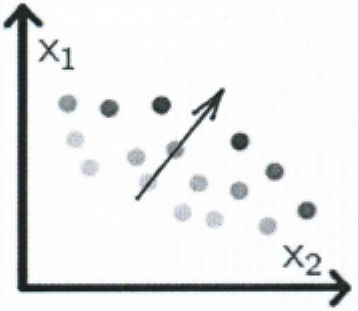


1(a) First PCA component

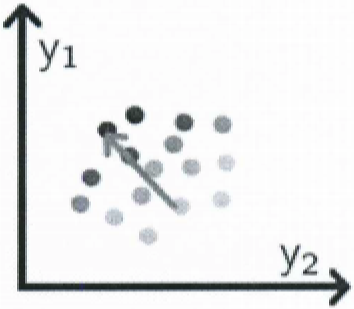


1(b) First LDA component

2. (2pts) In practice, each data point may have multiple vector-valued properties, e.g. a gene has its expression levels as well as the position on the genome. The goal of CCA is to reduce the dimensionality of the properties jointly. Suppose we have data points with two properties $\mathbf{x}$ and $\mathbf{y}$, each of which is a 2-dimension vector. This 4-dimensional data is shown in the pair of figures below; different data points are shown in different gray scales. CCA *finds* $(\mathbf{u}, \mathbf{v})$ *to maximize the correlation* $\widehat{corr}(\mathbf{u}^T\mathbf{x})(\mathbf{v}^T\mathbf{y})$. In figure 2(b) we have given the direction of vector $\mathbf{v}$, plot the vector $\mathbf{u}$ in figure 2(a).



2(a)



2(b)

1 [Points] Short Questions

The following short questions should be answered with at most two sentences, and/or a picture. For the **(true/false)** questions, answer true or false. If you answer true, provide a short justification, if false explain why or provide a small counterexample.

1. [points] Your billionaire friend needs your help. She needs to classify job applications into good/bad categories, and also to detect job applicants who lie in their applications using density estimation to detect outliers. To meet these needs, do you recommend using a discriminative or [✓ generative classifier]? Why?

   for density estimation,
   need p(x|y)

2. [points] Your billionaire friend also wants to classify software applications to detect bug-prone applications using features of the source code. This pilot project only has a few applications to be used as training data, though. To create the most accurate classifier, do you recommend using a [✓ discriminative] or generative classifier? Why?

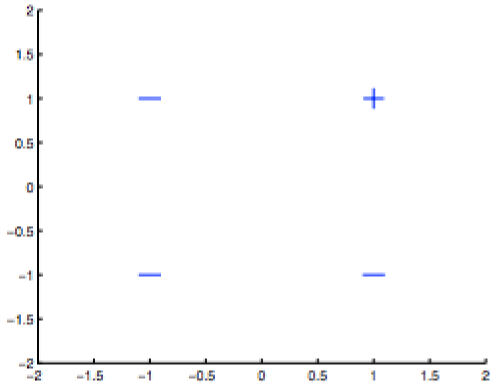
   based on
   convergence
   properties
   and some
   experimental
   observations

3. [points] Finally, your billionaire friend also wants to classify companies to decide which one to acquire. This project has lots of training data based on several decades of research. To create the most accurate classifier, do you recommend using a discriminative or [✓ generative] classifier? Why?

6 [30 points] Decision Tree and Ensemble Methods

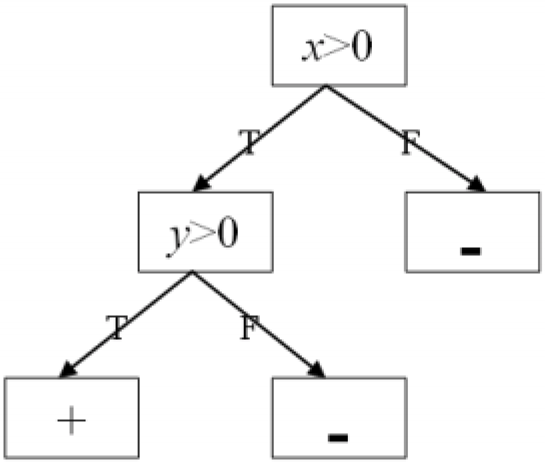
An ensemble classifier $H_T(x)$ is a collection of T weak classifiers $h_t(x)$, each with some weight α_t, $t = 1, \ldots, T$. Given a data point $x \in \mathbb{R}^d$, $H_T(x)$ predicts its label based on the weighted majority vote of the ensemble. In the binary case where the class label is either 1 or -1, $H_T(x) = \text{sgn}(\sum_{t=1}^{T} \alpha_t h_t(x))$, where $h_t(x) : \mathbb{R}^d \to \{-1, 1\}$, and $\text{sgn}(z) = 1$ if $z > 0$ and $\text{sgn}(z) = -1$ if $z \leq 0$. Boosting is an example of ensemble classifiers where the weights are calculated based on the training error of the weak classifier on the weighted training set.

1. [10 points] For the following data set,

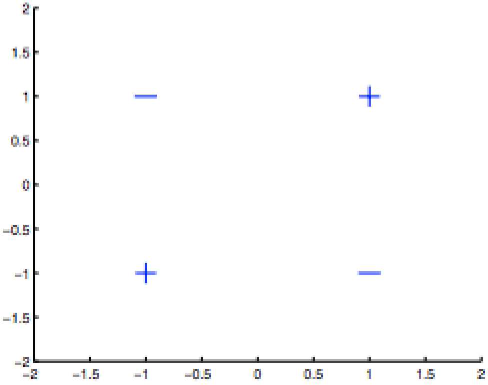


- Describe a binary decision tree with the minimum depth and consistent with the data;

★ **SOLUTION:** The decision tree is as follows.

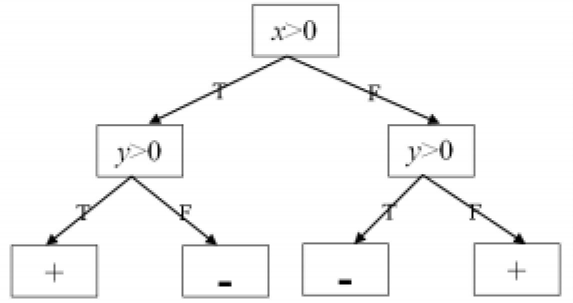


2. [10 points] For the following XOR data set,



- Describe a binary decision tree with the minimum depth and consistent with the data;

★ **SOLUTION:** The decision tree is as follows.

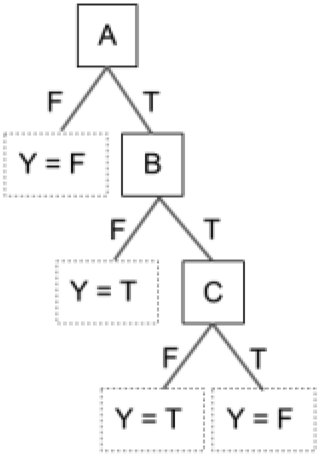


3. [6 points]

A	B	C	Y
F	F	F	F
T	F	T	T
T	T	F	T
T	T	T	F

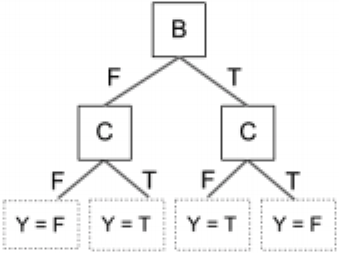
(a) [3 points] Using the dataset above, we want to build a decision tree which classifies Y as T/F given the binary variables A, B, C. Draw the tree that would be learned by the greedy algorithm with zero training error. You do not need to show any computation.

★ **ANSWER:**



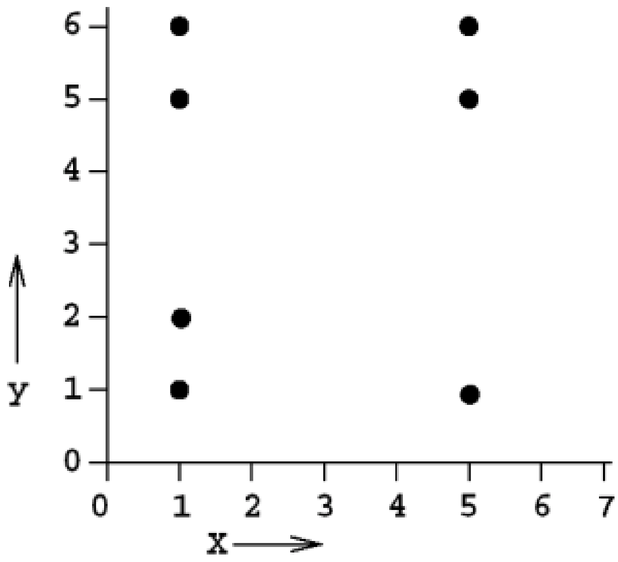
(b) [3 points] Is this tree optimal (i.e. does it get zero training error with minimal depth)? Explain in less than two sentences. If it is not optimal, draw the optimal tree as well.

★ **ANSWER:** Although we get a better information gain by first splitting on A, Y is just a function of B/C: $Y = B\text{xor}C$. Thus, we can build a tree of depth 2 which classifies correctly, and is optimal.



Problem 4. Instance based learning (8 points)

The following picture shows a dataset with one real-valued input x and one real-valued output y. There are seven training points.



Suppose you are training using kernel regression using some unspecified kernel function. The only thing you know about the kernel function is that it is a monotonically decreasing function of distance that decays to zero at a distance of 3 units (and is strictly greater than zero at a distance of less than 3 units).

(a) (*2 pts*) What is the predicted value of y when x = 1?

Answer: $\frac{1+2+5+6}{4} = 3.5$

(b) (*2 pts*) What is the predicted value of y when x = 3?

Answer: $\frac{1+2+5+6+1+5+6}{7} = 26/7$

(c) (*2 pts*) What is the predicted value of y when x = 4?

Answer: $\frac{1+5+6}{3} = 4$

(d) (*2 pts*) What is the predicted value of y when x = 7?

Answer: $\frac{1+5+6}{3} = 4$

2 Instance-Based Learning (7pts)

1. Consider the following training set in the 2-dimensional Euclidean space:

x	y	Class
-1	1	$-$
0	1	$+$
0	2	$-$
1	-1	$-$
1	0	$+$
1	2	$+$
2	2	$-$
2	3	$+$

Figure 1 shows a visualization of the data.

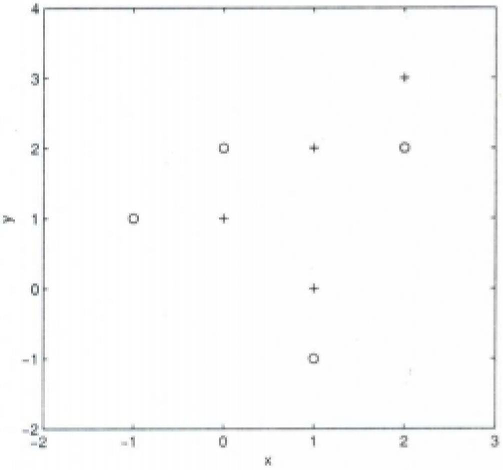


Figure 1: Dataset for Problem 2

(a) (1pt) What is the prediction of the 3-nearest-neighbor classifier at the point (1,1)?

+

(b) (1pt) What is the prediction of the 5-nearest-neighbor classifier at the point (1,1)?

+

(c) (1pt) What is the prediction of the 7-nearest-neighbor classifier at the point (1,1)?

−

7 SVM and Kernel Methods (8 points)

(a) Kernel functions implicitly define some mapping function $\phi(\cdot)$ that transforms an input instance $\mathbf{x} \in \mathbb{R}^d$ to a high dimensional feature space Q by giving the form of dot product in Q: $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$.

Assume we use radial basis kernel function $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{1}{2}\|\mathbf{x}_i - \mathbf{x}_j\|^2)$. Thus we assume that there's some implicit unknown function $\phi(\mathbf{x})$ such that

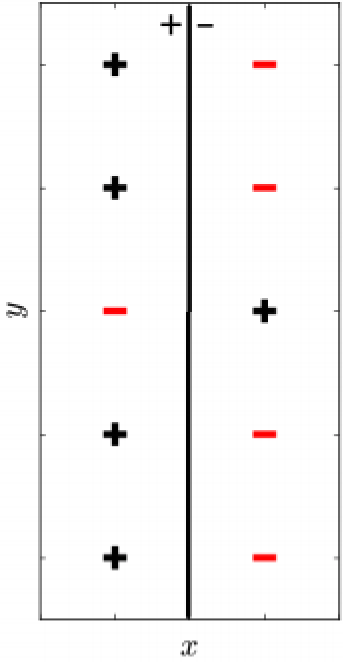
$$\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) = K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{1}{2}\|\mathbf{x}_i - \mathbf{x}_j\|^2)$$

Prove that for any two input instances $\mathbf{x}_i$ and $\mathbf{x}_j$, the squared Euclidean distance of their corresponding points in the feature space Q is less than 2, i.e. prove that $\|\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)\|^2 < 2$.

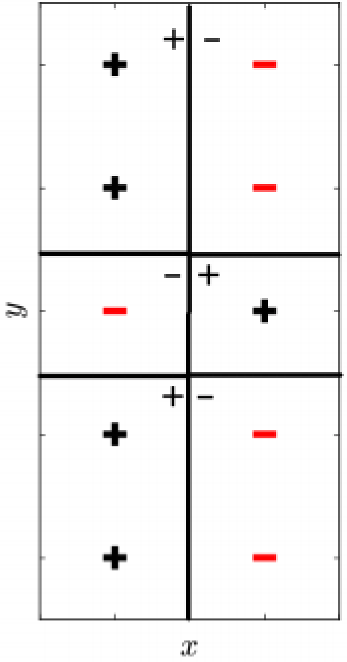
$$\begin{aligned}
&\|\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)\|^2 \\
=&(\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)) \cdot (\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)) \\
=&\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_i) + \phi(\mathbf{x}_j) \cdot \phi(\mathbf{x}_j) - 2 \cdot \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \\
=&2 - 2\exp(-\frac{1}{2}\|\mathbf{x}_i - \mathbf{x}_j\|^2) \\
<&2
\end{aligned}$$

1 [24 points] Short Answer

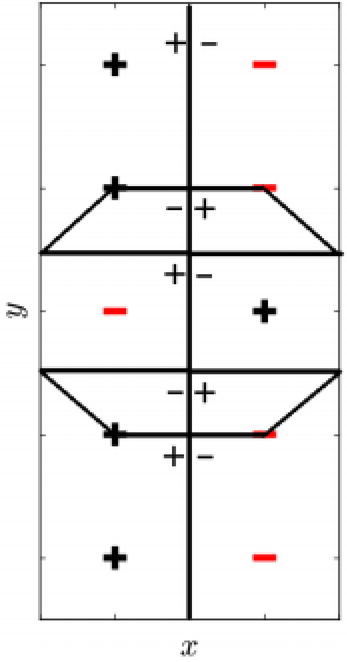
1. [6 points] On the 2D dataset below, draw the decision boundaries learned by the following algorithms (using the features x/y). **Be sure to mark which regions are labeled positive or negative, and assume that ties are broken arbitrarily.**



(a) Logistic regression ($\lambda = 0$)



(b) 1-NN



(c) 3-NN

9 Nearest neighbor and cross-validation

At some point during this question you may find it useful to use the fact that if U and V are two independent real-valued random variables then $Var[aU + bV] = a^2\, Var[U] + b^2\, Var[V]$.

Suppose you have 10,000 datapoints $\{(x_k, y_k) : k = 1, 2, ..., 10000\}$. Your dataset has one input and one output. The kth datapoint is generated by the following recipe:

$$\begin{aligned} x_k &= k/10000 \\ y_k &\sim N(0, 2^2) \end{aligned}$$

So that y_k is all noise: drawn from a Gaussian with mean 0 and variance $\sigma^2 = 4$ (and standard deviation $\sigma = 2$). Note that its value is independent of all the other y values. You are considering two learning algorithms:

- **Algorithm NN:** 1-nearest neighbor.
- **Algorithm Zero:** Always predict zero.

(a) What is the expected Mean Squared Training Error for **Algorithm NN**?

0

(b) What is the expected Mean Squared Training Error for **Algorithm Zero**?

4

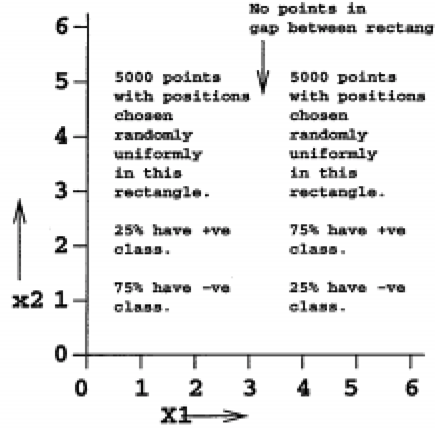
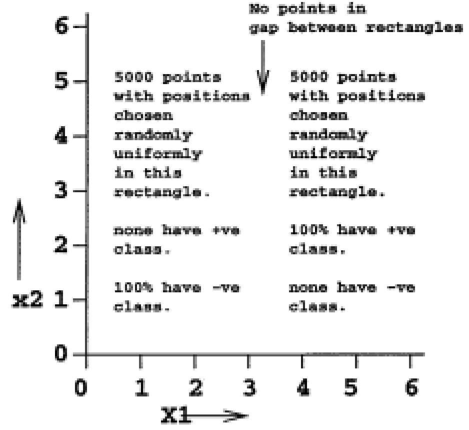
(c) What is the expected Mean Squared Leave-one-out Cross-validation Error for **Algorithm NN**?

8 = E[(xk - x[x+1])^2]

(d) What is the expected Mean Squared Leave-one-out Cross-validation Error for **Algorithm Zero**?

4

7 Nearest Neighbor and Cross-Validation

Recipe for making training set of 10,000 datapoints with two real-valued inputs and one binary output class:	Recipe for making test set of 10,000 datapoints with two real-valued inputs and one binary output class:
	

No points in gap between rectangles
5000 points with positions chosen randomly uniformly in this rectangle.
25% have +ve class.
75% have -ve class.
5000 points with positions chosen randomly uniformly in this rectangle.
75% have +ve class.
25% have -ve class.
x2
X1
0
1
2
3
4
5
6

No points in gap between rectangles
5000 points with positions chosen randomly uniformly in this rectangle.
none have +ve class.
100% have -ve class.
5000 points with positions chosen randomly uniformly in this rectangle.
100% have +ve class.
none have -ve class.
x2
X1
0
1
2
3
4
5
6

Using the above recipes for making training and test sets you will see that the training set is noisy: in either region, 25% of the data comes from the minority class. The test set is noise-free.

In each of the following questions, circle the answer that most closely defines the expected error rate, expressed as a fraction.

(a) What is the expected training set error using one-nearest-neighbor?

[✓ 0] 1/8 1/4 3/8 1/3 1/2 5/8 2/3 3/4 7/8 1

(b) What is the expected leave-one-out cross-validation error on the training set using one-nearest-neighbor? $\frac{1}{4} \times 3/4 + 3/4 \times \frac{1}{4} = 3/8$

0 1/8 1/4 [✓ 3/8] 1/3 1/2 5/8 2/3 3/4 7/8 1

(c) What is the expected test set error if we train on the training set, test on the test set, and use one-nearest-neighbor?

0 1/8 1/4 [✓ 3/8] 1/3 1/2 5/8 2/3 3/4 7/8 1

(d) What is the expected training set error using 21-nearest-neighbor?

0 1/8 [✓ 1/4] 3/8 1/3 1/2 5/8 2/3 3/4 7/8 1

(e) What is the expected leave-one-out cross-validation error on the training set using 21-nearest-neighbor?

0 1/8 [✓ 1/4] 3/8 1/3 1/2 5/8 2/3 3/4 7/8 1

(f) What is the expected test set error if we train on the training set, test on the test set, and use 21-nearest-neighbor?

0 1/8 [✓ 1/4] 3/8 1/3 1/2 5/8 2/3 3/4 7/8 1

1 Loss, Regularization and Optimization [10 points]

1.1 Quick Questions [4 points]

Explain in **one or two sentences** why the statements are true (or false).

1. L2 loss is more robust to outliers than L1 loss.

   Solution:
   False
   The gradient of L2 loss can grow without bound whereas the L1 loss gradient is bounded, hence the influence of an outlier is limited.

2. Logistic loss is better than L2 loss in classification tasks.

   Solution:
   True
   With logistic loss, correctly classified points that are far away from the decision boundary have much less impact on the decision boundary

3. In terms of feature selection, L2 regularization is preferred since it comes up with sparse solutions.

   Solution:
   False
   L1 regularization (LASSO) comes up with sparse solutions due to nonvanishing gradient at 0.

(a) Describe the difference between a *maximum likelihood* hypothesis and a *maximum a posteriori* hypothesis.

MLE = maximize P(data | parameters) by searching
over parameters

MAP = Maximize P(parameters | data) by searching
over params, and accounting for prior
over params

2. [6 points] A random variable follows an *exponential* distribution with parameter λ ($\lambda > 0$) if it has the following density:

$$p(t) = \lambda e^{-\lambda t}, \quad t \in [0, \infty)$$

This distribution is often used to model waiting times between events. Imagine you are given i.i.d. data $T = (t_1, \dots, t_n)$ where each t_i is modeled as being drawn from an exponential distribution with parameter λ.

(a) [3 points] Compute the log-probability of T given λ. (Turn products into sums when possible).

★ **ANSWER:**

$$\begin{aligned}
\ln p(T) &= \ln \prod_i p(t_i) \\
\ln p(T) &= \sum_i \ln(\lambda e^{-\lambda t_i}) \\
\ln p(T) &= \sum_i \ln \lambda - \lambda t_i \\
\ln p(T) &= \boxed{n \ln \lambda - \lambda \sum_i t_i}
\end{aligned}$$

(b) [3 points] Solve for $\hat{\lambda}_{MLE}$.

★ **ANSWER:**

$$\begin{aligned}
\frac{\partial}{\partial \lambda}(n \ln \lambda - \lambda \sum_i t_i) &= 0 \\
\frac{n}{\lambda} - \sum_i t_i &= 0 \\
\hat{\lambda}_{MLE} &= \boxed{\frac{n}{\sum_i t_i}}
\end{aligned}$$

5 Bayes Rule (19 points)

(a) *(4 points)* I give you the following fact:

$$P(A|B) = 2/3$$

Do you have enough information to compute $P(B|A)$? If not, write "not enough info". If so, compute the value of $P(B|A)$.

Not Enough Info

(b) *(5 points)* Instead, I give you the following facts:

$$\begin{aligned} P(A|B) &= 2/3 \\ P(A|{\sim}B) &= 1/3 \end{aligned}$$

Do you now have enough information to compute $P(B|A)$? If not, write "not enough info". If so, compute the value of $P(B|A)$.

Not enough Info

(c) *(5 points)* Instead, I give you the following facts:

$$\begin{aligned} P(A|B) &= 2/3 \\ P(A|{\sim}B) &= 1/3 \\ P(B) &= 1/3 \end{aligned}$$

Do you now have enough information to compute $P(B|A)$? If not, write "not enough info". If so, compute the value of $P(B|A)$.

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|{\sim}B)P({\sim}B)} = \frac{\frac{2}{3} \times \frac{1}{3}}{\frac{2}{3} \times \frac{1}{3} + \frac{1}{3} \times \frac{2}{3}} = \frac{1}{2}$$

(d) *(5 points)* Instead, I give you the following facts:

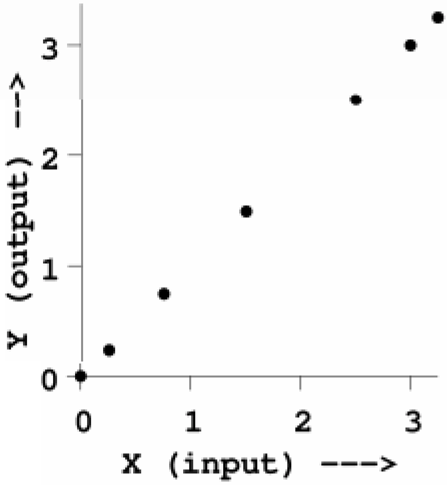
$$\begin{aligned} P(A|B) &= 2/3 \\ P(A|{\sim}B) &= 1/3 \\ P(B) &= 1/3 \\ P(A) &= 4/9 \end{aligned}$$

Do you now have enough information to compute $P(B|A)$? If not, write "not enough info". If so, compute the value of $P(B|A)$.

Still $\frac{1}{2}$ (of course)

3 Regression

(a) Consider the following data with one input and one output.



- (i) What is the mean squared training set error of running linear regression on this data (using the model $y = w_0 + w_1 x$)?

  0

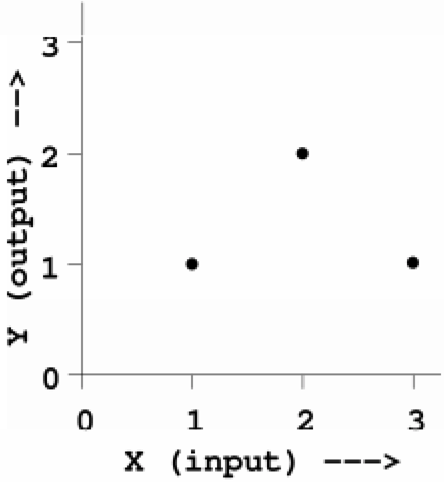
- (ii) What is the mean squared test set error of running linear regression on this data, assuming the rightmost three points are in the test set, and the others are in the training set.

  0

- (iii) What is the mean squared leave-one-out cross-validation (LOOCV) error of running linear regression on this data?

0

(b) Consider the following data with one input and one output.



X	Y
1	1
2	2
3	1

- (i) What is the mean squared training set error of running linear regression on this data (using the model $y = w_0 + w_1 x$)? (Hint: by symmetry it is clear that the best fit to the three datapoints is a horizontal line).

```
SSE = (1/3)^2 + (2/3)^2 + (1/3)^2 = 6/9
MSE = SSE/3 = 2/9
```

- (ii) What is the mean squared leave-one-out cross-validation (LOOCV) error of running linear regression on this data?

```
1/3 * (2^2 + 1^2 + 2^2) = 9/3 = 3
```