

Exercises on text classification

Ion Androutsopoulos, 2017–18 (last revision 22/2/18)

Submit as a group of 3–4 members a report for exercise 17 (max. 5 pages, PDF format). Include in your report all the required information, especially experimental results. Do not include code in the report, but include a link to a shared folder or repository (e.g. in Dropbox, GitHub, Bitbucket) containing your code.

1. Let $\langle X_1, X_2, X_3 \rangle = \langle 0, 1, 0 \rangle$ be the feature vector of an unseen object to be classified. We use a k -NN classifier, with two classes ($C = 0, C = 1$), $k = 3$, and the number of different features (slide 17) as the distance measure. The training dataset contains the four vectors (rows) of the table (excluding the first column). Which class will the unseen object be classified in?

d	0	1	0	$C?$
2	1	0	0	0
1	0	1	1	0
3	1	0	1	1
2	1	1	1	1

Answer: The first column of the table shows the distance (d) from the unseen object to the corresponding training instances. The $k = 3$ nearest neighbors are at distances 1 and 2. The majority class among the k nearest neighbors is $C = 0$. Hence, the unseen object will be classified in $C = 0$.

2. Two training examples are *inconsistent*, if they have the same feature values, but different “correct” class labels. If there are *no* inconsistent training examples and we evaluate a k -NN classifier on the same dataset we used for training, then its accuracy will be:

- 100%
 100% if $k = 1$, but may not be 100% if $k \neq 1$
 none of the above (the accuracy depends on the dataset).

Answer: If $k = 1$, every test instance will be classified in the class of its closest training instance, and the closest training instance will be itself (or a copy of itself, with the same correct class label, since there are no inconsistencies), given that the test dataset is the same as the training dataset. Hence, for $k = 1$ every test instance will be classified correctly and the accuracy will be 100%.

If $k \neq 1$, each test instance will be classified in the majority class of its k nearest training (and test) instances, and the majority class may not be the correct one. Hence, for $k \neq 1$ the accuracy may not be 100%.

3. (a) Based on the training examples of the table, the entropy of C is:

$H(C) = 1$ $H(C) = 0$ $H(C) = 1/2$

Answer: Based on the training examples, $P(C=\text{positive}) = P(C=\text{negative}) = 1/2$. Hence, C has two equally probable values and, therefore, its entropy is maximum. For two possible outcomes, the maximum entropy value is 1. You should be able to reach the same conclusion by computing the entropy of C using the definition of entropy.

X	Y	Z	C
0	1	0	positive
0	1	1	positive
0	1	0	positive
1	0	1	positive
1	1	0	negative
0	0	1	negative
0	0	0	negative
0	0	1	negative

(b) Based on the training examples of the table, the highest information gain score is:

$IG(C, X)$ $IG(C, Y)$ $IG(C, Z)$

Answer: The training data show that if we learn that $Y = 1$, it is very likely (probability $3/4$) that $C = \text{positive}$; and if we learn that $Y = 0$, it is very likely (probability $3/4$) that $C =$

negative. Hence, knowing the value of Y reduces the entropy of C , which was initially maximum. You should be able to reach the same conclusion by computing $IG(C, Y)$.

By contrast, the training data show that if we learn that $X = 1$, the probability of $C = \text{positive}$ remains $\frac{1}{2}$, equal to the probability of $C = \text{negative}$; and if we learn that $X = 0$, the probability of $C = \text{positive}$ remains $\frac{1}{2}$, equal to the probability of $C = \text{negative}$. Hence, whatever the value of X , learning X still leaves us with two equally probable outcomes $C = \text{positive}$ and $C = \text{negative}$ and a maximum entropy, i.e., knowing X does not reduce the entropy of C , which implies that $IG(C, X) = 0$. You should be able to reach the same conclusion by computing $IG(C, X)$.

Similarly, $IG(C, Z) = 0$.

4. Confirm that $IG(X; C) = IG(C; X)$ (slide 9).

5. (a) Explain why supervised machine learning algorithms must be evaluated on different data than those used for training.

Answer: If evaluated on the training dataset, an algorithm that simply memorizes the training instances and their correct class labels would obtain an accuracy score of 100%, without this score being indicative of how well it would perform on unseen instances. More generally, a supervised learning algorithm often performs better (due to overfitting) on the training dataset than on unseen data.

(b) A researcher submitted a conference paper that described a named entity recognizer (NER) that used supervised machine learning. The paper reported experiments with different feature sets. For each feature set, the NER had been trained on a training corpus (the same for all feature sets) and had been tested on a separate test corpus (the same for all feature sets). The paper listed test F1 results for each feature set, which were used to select the best feature set. With the best feature set, the NER of the paper had better F1 (and other) scores on the test corpus, compared to other NER systems that had been trained and evaluated on the same training and test corpora, respectively. However, the reviewers rejected the paper, noting that the test results were unreliable and that the comparison to the other systems was wrong, because the test set was in effect also used during training. Were the reviewers right or not? Explain why. If they were right, what should the researcher do to address the problem?

Answer: The reviewers were right to complain, because the researcher selected the feature set that led to the best results on the test dataset, i.e., the test dataset was used for feature selection, which in effect is part of training (or at least configuring) the system. The chosen feature set may contain features that work well for the particular test set (especially if the test set is small), but may not work that well on other test sets; in effect, the system may have overfitted the particular test set. The researcher should select the feature set that leads to the best results on a held-out part of the training set (or perform a cross-validation on the training set to select the feature set), and then evaluate the NER (with the chosen feature set) on the test set.

6. (a) Let $\langle X_1, X_2, X_3 \rangle = \langle 0, 1, 0 \rangle$ be the feature vector of a text to be classified. We use a multivariate Bernoulli Naive Bayes classifier, with two classes ($C = 0, C = 1$), and the training data of the table. What will be the decision of the classifier? Use Laplace estimates for $P(X_i|C)$. All the features X_i are Boolean.

X_1	X_2	X_3	C
1	0	0	0
0	1	1	0
1	0	1	1
1	1	1	1

Answer:

$$P(C = 1 | X_1 = 0, X_2 = 1, X_3 = 0) \cong$$

$$\begin{aligned} & \frac{P(C = 1)}{P(X_1 = 0, X_2 = 1, X_3 = 0)} \cdot P(X_1 = 0|C = 1) \cdot P(X_2 = 1|C = 1) \cdot P(X_3 = 0|C = 1) \\ & \cong \frac{1/2}{P(X_1 = 0, X_2 = 1, X_3 = 0)} \cdot \frac{0+1}{2+2} \cdot \frac{1+1}{2+2} \cdot \frac{0+1}{2+2} \\ & = \frac{1/2}{P(X_1 = 0, X_2 = 1, X_3 = 0)} \cdot \frac{1}{4} \cdot \frac{2}{4} \cdot \frac{1}{4} \end{aligned}$$

$$\begin{aligned} P(C = 0|X_1 = 0, X_2 = 1, X_3 = 0) & \cong \\ & \frac{P(C = 0)}{P(X_1 = 0, X_2 = 1, X_3 = 0)} \cdot P(X_1 = 0|C = 0) \cdot P(X_2 = 1|C = 0) \cdot P(X_3 = 0|C = 0) \\ & \cong \frac{1/2}{P(X_1 = 0, X_2 = 1, X_3 = 0)} \cdot \frac{1+1}{2+2} \cdot \frac{1+1}{2+2} \cdot \frac{1+1}{2+2} \\ & = \frac{1/2}{P(X_1 = 0, X_2 = 1, X_3 = 0)} \cdot \frac{2}{4} \cdot \frac{2}{4} \cdot \frac{2}{4} \end{aligned}$$

Hence, the decision will be $C = 0$.

7. Let $\langle X_1, X_2, X_3, X_4 \rangle = \langle b, d, b, a \rangle$ be the feature vector of an object to be classified. We use a multivariate Bernoulli Naive Bayes classifier, with three classes ($C = 0, C = 1, C = 3$), and the training data of the table. What will be the decision of the classifier? Use Laplace estimates for $P(X_i|C)$. Each feature X_i has four possible values: a, b, c, d.

X_1	X_2	X_3	X_4	C
a	b	b	a	1
b	a	a	b	1
b	b	a	b	1
a	a	b	b	2
a	b	b	b	2
b	a	b	a	2
c	d	d	c	3
d	c	c	d	3
d	d	c	d	3

Answer: From the training data, we see that $P(C = 1) = P(C = 2) = P(C = 3) = 3/9$, i.e., the priori probabilities do not influence the decision.

Using Laplace estimates, we obtain:

$$\begin{aligned} P(X_1 = b|C = 1) &= \frac{2+1}{3+4} = \frac{3}{7} & P(X_2 = d|C = 1) &= \frac{0+1}{3+4} = \frac{1}{7} \\ P(X_3 = b|C = 1) &= \frac{1+1}{3+4} = \frac{2}{7} & P(X_4 = a|C = 1) &= \frac{1+1}{3+4} = \frac{2}{7} \end{aligned}$$

and:

$$\begin{aligned} P(X_1 = b|C = 2) &= \frac{1+1}{3+4} = \frac{2}{7} & P(X_2 = d|C = 2) &= \frac{0+1}{3+4} = \frac{1}{7} \\ P(X_3 = b|C = 2) &= \frac{3+1}{3+4} = \frac{4}{7} & P(X_4 = a|C = 2) &= \frac{1+1}{3+4} = \frac{2}{7} \end{aligned}$$

and:

$$\begin{aligned} P(X_1 = b|C = 3) &= \frac{0+1}{3+4} = \frac{1}{7} & P(X_2 = d|C = 3) &= \frac{2+1}{3+4} = \frac{3}{7} \\ P(X_3 = b|C = 3) &= \frac{0+1}{3+4} = \frac{1}{7} & P(X_4 = a|C = 3) &= \frac{0+1}{3+4} = \frac{1}{7} \end{aligned}$$

Hence:

$$P(X_1 = b|C = 1) \cdot P(X_2 = d|C = 1) \cdot P(X_3 = b|C = 1) \cdot P(X_4 = a|C = 1) = \frac{3 \cdot 1 \cdot 2 \cdot 2}{7 \cdot 7 \cdot 7 \cdot 7}$$

$$P(X_1 = b|C = 2) \cdot P(X_2 = d|C = 2) \cdot P(X_3 = b|C = 2) \cdot P(X_4 = a|C = 2) = \frac{2 \cdot 1 \cdot 4 \cdot 2}{7 \cdot 7 \cdot 7 \cdot 7}$$

$$P(X_1 = b|C = 3) \cdot P(X_2 = d|C = 3) \cdot P(X_3 = b|C = 4) \cdot P(X_4 = a|C = 3) = \frac{1 \cdot 3 \cdot 1 \cdot 1}{7 \cdot 7 \cdot 7 \cdot 7}$$

Therefore, the decision is $C = 2$.

8. We use a multinomial Naive Bayes classifier, with two classes ($C = 0$, $C = 1$) and m Boolean (binary) features X_1, \dots, X_m , which decides $C = 1$ if:

$$P(C = 1) \cdot \prod_{i=1}^m P(X_i = x_i|C = 1) \geq K$$

and $C = 0$ otherwise. K is a constant. Show that this classifier is a linear separator. Hint: If we represent by t_i the outcome denoted by $X_i = 1$ (e.g., the occurrence of a particular word), then:

$$P(X_i = x_i|C = 1) = P(t_i|C = 1)^{x_i} \cdot [1 - P(t_i|C = 1)]^{1-x_i}$$

Recall, also, that $\log(a \cdot b) = \log a + \log b$ and $\log a^b = b \cdot \log a$.

Answer: The classifier decides $C = 1$ if and only if (iff):

$$\log[P(C = 1) \cdot \prod_{i=1}^m P(X_i = x_i|C = 1)] \geq \log K$$

which is equivalent to:

$$\log P(C = 1) + \sum_{i=1}^m \log P(X_i = x_i|C = 1) \geq \log K$$

and:

$$\log P(C = 1) + \sum_{i=1}^m \log\{P(t_i|C = 1)^{x_i} \cdot [1 - P(t_i|C = 1)]^{1-x_i}\} \geq \log K$$

and:

$$\log P(C = 1) + \sum_{i=1}^m x_i \cdot \log P(t_i|C = 1) + (1 - x_i) \cdot \log[1 - P(t_i|C = 1)] \geq \log K$$

Setting $K_1 = \log P(C = 1)$, $K_{2,i} = \log P(t_i|C = 1)$, $K_{3,i} = \log P[1 - P(t_i|C = 1)]$, $K_4 = \log K$, the previous criterion becomes:

$$K_1 + \sum_{i=1}^m [x_i \cdot K_{2,i} + (1 - x_i) \cdot K_{3,i}] \geq K_4$$

which is equivalent to:

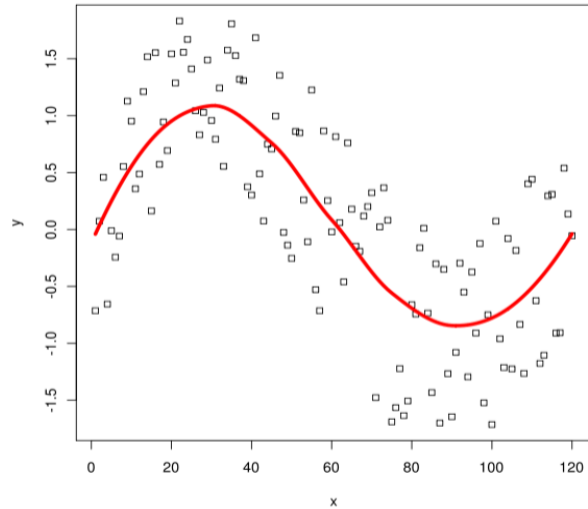
$$K_1 + \sum_{i=1}^m K_{3,i} - K_4 + \sum_{i=1}^m x_i \cdot (K_{2,i} - K_{3,i}) \geq 0$$

Setting $w_0 = K_1 + \sum_{i=1}^m K_{3,i} - K_4$ and $w_i = (K_{2,i} - K_{3,i})$, the criterion for $C = 1$ becomes:

$$w_0 + \sum_{i=1}^m w_i \cdot x_i \geq 0$$

Hence, the classifier is a linear separator.

9. The dots of the figure are a sample from a population that actually follows the function $y = f(x)$, whose curve is the solid line.¹ Due to noise (e.g., introduced by the measurements), the dots are not exactly on the solid line. We wish to learn from the sample a function $y = h(x)$ that will approximate the (unknown to us) true $f(x)$.



(a) Explain why linear regression would not lead to a good $h(x)$.

Answer: Linear regression learns straight lines (more generally, planes or hyperplanes). In our case, the target curve $y = f(x)$ cannot be approximated well by a single straight line.

(b) How could we learn a better $h(x)$ using k -NN regression (or variant)? What would we do during training? After training, given an x , how would we obtain $y = h(x)$;

Answer: During training, we would simply store all the coordinates (x, y) of the instances (dots) of the sample. After training, given an x' , we would retrieve the k (training) instances of the sample whose x values are closest to x' and we would return their average y . We could also assign weights to the y values of the k neighbors (e.g., inversely proportional to the distance from the neighbor's x to x').

10. Derive the weights update rule of least squares linear regression, when using stochastic gradient descent.

Answer: With stochastic gradient descent, the weights of least squares linear regression are updated as follows:

$$\vec{w} \leftarrow \vec{w} - \eta \cdot \nabla_{\vec{w}} E_i(\vec{w}), \text{ where } E_i(\vec{w}) = \frac{1}{2} [f_{\vec{w}}(\vec{x}^{(i)}) - y^{(i)}]^2$$

and:

$$\nabla_{\vec{w}} E_i(\vec{w}) = \left\langle \frac{\partial E_i(\vec{w})}{\partial w_0}, \frac{\partial E_i(\vec{w})}{\partial w_1}, \dots, \frac{\partial E_i(\vec{w})}{\partial w_l}, \dots, \frac{\partial E_i(\vec{w})}{\partial w_n} \right\rangle$$

For $l \in \{0, \dots, n\}$:

$$\frac{\partial E_i(\vec{w})}{\partial w_l} = [f_{\vec{w}}(\vec{x}^{(i)}) - y^{(i)}] \cdot x_l^{(i)}$$

Hence:

$$\begin{aligned} \nabla_{\vec{w}} E_i(\vec{w}) &= [f_{\vec{w}}(\vec{x}^{(i)}) - y^{(i)}] \cdot \langle x_1^{(i)}, \dots, x_l^{(i)}, \dots, x_n^{(i)} \rangle = \\ &= [f_{\vec{w}}(\vec{x}^{(i)}) - y^{(i)}] \cdot \vec{x}^{(i)} \end{aligned}$$

and the weights update rule is:

$$\vec{w} \leftarrow \vec{w} - \eta \cdot [f_{\vec{w}}(\vec{x}^{(i)}) - y^{(i)}] \cdot \vec{x}^{(i)}$$

11. Derive the weights update rule of logistic regression, when using batch gradient ascent.

¹ Figure from http://en.wikipedia.org/wiki/Local_regression.

Answer: With batch gradient ascent, the weights of logistic regression are updated as follows (using natural logarithms):

$$\vec{w} \leftarrow \vec{w} + \eta \cdot \nabla_{\vec{w}} l(\vec{w})$$

where:

$$l(\vec{w}) = \sum_{i=1}^m y^{(i)} \ln P(c_+ | \vec{x}^{(i)}; \vec{w}) + (1 - y^{(i)}) \ln P(c_- | \vec{x}^{(i)}; \vec{w}) \quad (1)$$

and:

$$P(c_+ | \vec{x}; \vec{w}) = 1 / (1 + e^{-\vec{w}\vec{x}}) \quad (2)$$

$$P(c_- | \vec{x}; \vec{w}) = e^{-\vec{w}\vec{x}} / (1 + e^{-\vec{w}\vec{x}}) \quad (3)$$

Using (2) and (3) in (1), we obtain:

$$\begin{aligned} l(\vec{w}) &= \sum_{i=1}^m y^{(i)} \ln \left(\frac{1}{1 + e^{-\vec{w}\vec{x}^{(i)}}} \right) + (1 - y^{(i)}) \ln \left(\frac{e^{-\vec{w}\vec{x}^{(i)}}}{1 + e^{-\vec{w}\vec{x}^{(i)}}} \right) = \\ &= \sum_{i=1}^m -y^{(i)} \ln(1 + e^{-\vec{w}\vec{x}^{(i)}}) + \ln(e^{-\vec{w}\vec{x}^{(i)}}) - \ln(1 + e^{-\vec{w}\vec{x}^{(i)}}) - y^{(i)} \ln(e^{-\vec{w}\vec{x}^{(i)}}) \\ &\quad + y^{(i)} \ln(1 + e^{-\vec{w}\vec{x}^{(i)}}) \\ &= \sum_{i=1}^m \ln(e^{-\vec{w}\vec{x}^{(i)}}) - \ln(1 + e^{-\vec{w}\vec{x}^{(i)}}) - y^{(i)} \ln(e^{-\vec{w}\vec{x}^{(i)}}) \end{aligned}$$

The gradient is:

$$\nabla_{\vec{w}} l(\vec{w}) = \left\langle \frac{\partial l(\vec{w})}{\partial w_0}, \frac{\partial l(\vec{w})}{\partial w_1}, \dots, \frac{\partial l(\vec{w})}{\partial w_l}, \dots, \frac{\partial l(\vec{w})}{\partial w_n} \right\rangle$$

and for $l \in \{0, \dots, n\}$:

$$\begin{aligned} \frac{\partial l(\vec{w})}{\partial w_l} &= - \sum_{i=1}^m \frac{e^{-\vec{w}\vec{x}^{(i)}} \cdot x_l^{(i)}}{e^{-\vec{w}\vec{x}^{(i)}}} - \frac{e^{-\vec{w}\vec{x}^{(i)}} \cdot x_l^{(i)}}{1 + e^{-\vec{w}\vec{x}^{(i)}}} - y^{(i)} \frac{e^{-\vec{w}\vec{x}^{(i)}} \cdot x_l^{(i)}}{e^{-\vec{w}\vec{x}^{(i)}}} \\ &= - \sum_{i=1}^m x_l^{(i)} - x_l^{(i)} \frac{e^{-\vec{w}\vec{x}^{(i)}}}{1 + e^{-\vec{w}\vec{x}^{(i)}}} - y^{(i)} x_l^{(i)} \\ &= - \sum_{i=1}^m \left(\frac{1 + e^{-\vec{w}\vec{x}^{(i)}} - e^{-\vec{w}\vec{x}^{(i)}}}{1 + e^{-\vec{w}\vec{x}^{(i)}}} - y^{(i)} \right) x_l^{(i)} \\ &= \sum_{i=1}^m [y^{(i)} - P(c_+ | \vec{x}^{(i)})] x_l^{(i)} \end{aligned}$$

Hence, the weights update rule is:

$$w_l \leftarrow w_l + \eta \cdot \sum_{i=1}^m [y^{(i)} - P(c_+ | \vec{x}^{(i)})] x_l^{(i)}$$

12. We trained a logistic regression classifier on the training instances (dots) of the figure. There are two classes (black, white) and two features (x_1, x_2). If we evaluate the classifier on the training (same) instances, will it classify all the training instances correctly? If yes, why? If not, what can we do to help the classifier classify correctly all the training instances?

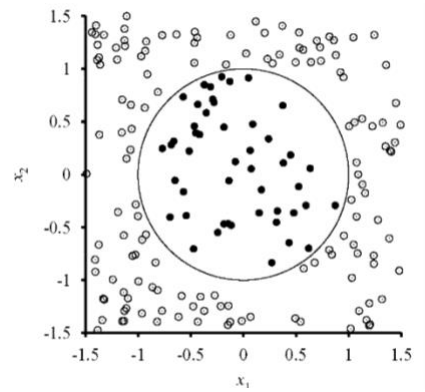


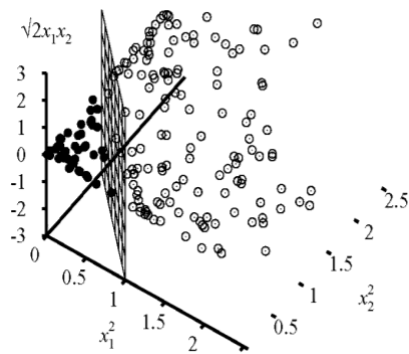
Figure from the book of Russel and Norvig.

Answer: Logistic regression classifiers are linear separators. They learn a straight line (in our case) or a hyperplane (more generally), and then classify unseen instances by computing whether or not they fall above or below the straight line (or hyperplane). The instances (dots) of the figure are not linearly separable (no straight line separates all the black from all the white dots). Hence, a logistic regression classifier cannot learn to separate them all (classify them all correctly), even if it is trained on the same instances.

By using more features, however, the instances may become linearly separable. For example, with the following transformation:

$$\vec{F}(\vec{x}) = \langle x_1^2, x_2^2, \sqrt{2}x_1x_2 \rangle$$

the instances are moved to a new, 3-dimensional vector space where they are linearly separable (2nd figure, also from the book of Russel & Norvig). Hence, a logistic regression classifier could learn to classify them all correctly, provided that we apply the transformation above.



13. A student trained a logistic regression classifier with (batch) gradient ascent. To speed up the training, the student used a large constant η value in the weights update rule, hoping that this way the gradient ascent would need fewer steps. However, the gradient ascent no longer converged, but seemed to be oscillating between two values. Why did this happen?

Answer: Because of the large η value, the gradient ascent possibly went over the maximum (overshot it), then made another (large) step towards the opposite direction (overshooting again the maximum) etc. Reducing the value of η at each step would probably have helped in this case.

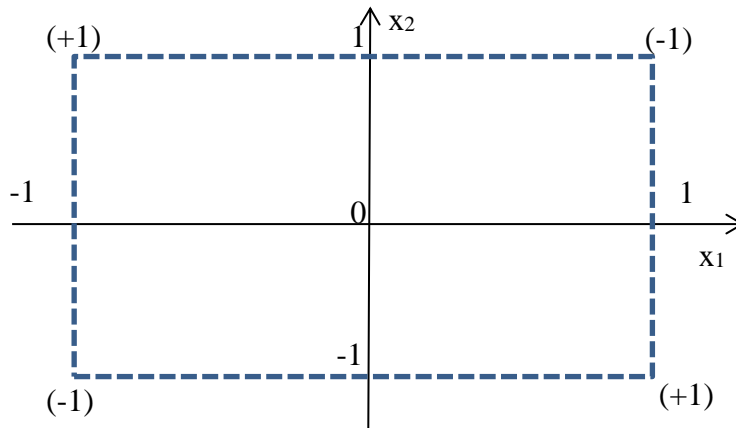
14. Confirm that maximizing the conditional-log likelihood of a logistic regression classifier is the same as minimizing its cross-entropy for all the training examples (slide 57).

15. [Optional] Show that an SVM can compute the XOR function. Use 1, -1 to represent *true* and *false*, respectively. For example, given the input $\langle -1, 1 \rangle$, the output should be 1; and given $\langle -1, -1 \rangle$, the output should be -1. Assume that the SVM transforms each input $\vec{x} = \langle x_1, x_2 \rangle$ of the original vector space to $\vec{F}(\vec{x}) = \langle x_1, x_1x_2 \rangle$ (in the new vector space). Plot the points $\langle x_1, x_1x_2 \rangle$ of the new vector space that correspond to all the possible inputs $\langle x_1, x_2 \rangle$ of XOR. Draw the separator (straight line) that the SVM would learn in the new vector space. What is the margin in the new vector space? Draw the corresponding separating lines in the original vector space. (Simplified form of exercise 20.12 of Russel & Norvig.)

Answer: The truth table of XOR is the following.

x_1	x_2	$XOR(x_1, x_2)$
-1	-1	-1
-1	1	1
1	-1	1
1	1	-1

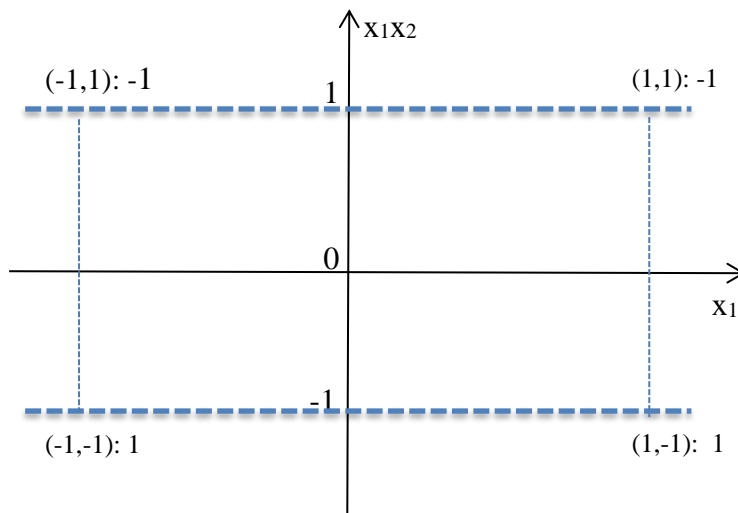
We wish to separate the two input vectors $\langle x_1, x_2 \rangle$ with positive $XOR(x_1, x_2)$ from the two input vectors with negative $XOR(x_1, x_2)$. In the original vector space, the four vectors are not linearly separable, as shown below:



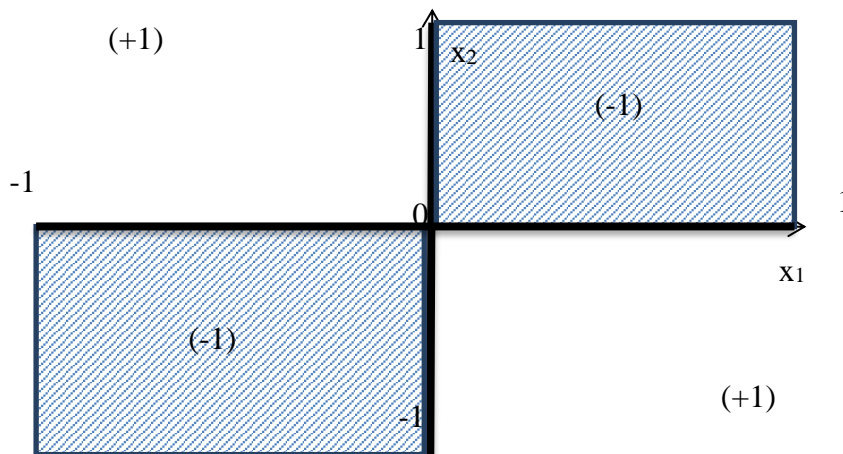
We extend the truth table to include the new feature $x_1 x_2$ of $\vec{F}(\vec{x}) = \langle x_1, x_1 x_2 \rangle$:

x_1	x_2	$x_1 x_2$	$XOR(x_1, x_2)$
-1	-1	1	-1
-1	1	-1	1
1	-1	-1	1
1	1	1	-1

The transformed four vectors $\langle x_1, x_1 x_2 \rangle$ are linearly separable in the new vector space, as shown below. The transformed vectors with negative $XOR(x_1, x_2)$ are now on the half-plane $x_1 x_2 \geq 1$ and the transformed vectors with positive $XOR(x_1, x_2)$ are now on the half-plane $x_1 x_2 \leq -1$. The maximum margin linear separator is the straight line $x_1 x_2 = 0$ and its margin is 2.



The half-plane $x_1x_2 \geq 0$ (negative class) of the new vector space corresponds to the shaded areas of the original vector space. The half-plane $x_1x_2 \leq 0$ (positive class) of the new vector space corresponds to the rest of the original vector space.



16. [Optional] Confirm the results of the hidden calculations (“...”) of slides 79 and 80 (kernel examples).

17. Develop a text classifier for a kind of texts of your choice (e.g., e-mail messages, tweets, customer reviews) and at least two classes (e.g., spam/ham, positive/negative/neutral).² You may use Boolean, TF, or TF-IDF features corresponding to words or n -grams, to which you can also add other features (e.g., length of the text). You may apply any feature selection (or dimensionality reduction) method you consider appropriate. You may also want to try using centroids of pre-trained word embeddings (slide 36).³ You can write your own code to produce feature vectors, perform feature selection (or dimensionality reduction) and train the

² For e-mail spam filtering, you may want to use the Ling-Spam or Enron-Spam datasets (available from <http://nlp.cs.aueb.gr/software.html>). For tweets, you may want to use datasets from <http://alt.qcri.org/semeval2016/task4/>. For customer reviews, you may want to use datasets from <http://alt.qcri.org/semeval2016/task5/>. Consult the instructor for further details.

³ Pre-trained word embeddings are available, for example, from <http://nlp.stanford.edu/projects/glove/>. See also word2vec (<https://code.google.com/archive/p/word2vec/>).

classifier (e.g., using SGD and the tricks of slides 59–60, in the case of logistic regression), or you can use existing implementations and software libraries.⁴ You should experiment with at least logistic regression, and optionally other learning algorithms (e.g., Naive Bayes, k -NN, SVM). Draw learning curves (slides 66, 69) with appropriate measures (e.g., accuracy, F1) and precision-recall curves (slide 24). Include experimental results of appropriate baselines (e.g., majority classifiers). Make sure that you use separate training and test data. Tune the feature set and hyper-parameters (e.g., regularization weight λ) on a held-out part of the training data or using a cross-validation (slide 26) on the training data. Document clearly in a short report (max. 5 pages) how your system works (e.g., what algorithms it uses, examples of input/output) and its experimental results (e.g., learning curves, precision-recall curves).

⁴ For example, LIBSVM (<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>), scikit-learn (<http://scikit-learn.org/stable/>), SVMlight (<http://svmlight.joachims.org/>), Weka (<http://www.cs.waikato.ac.nz/ml/weka/>).