10-601 Machine Learning, Fall 2011: Homework 5

Machine Learning Department
Carnegie Mellon University

Due: ???

Instructions There are **3** questions on this assignment. Please submit your completed homework to Sharon Cavlovich (GHC 8215) by **???**. Submit your homework as **3 separate** sets of pages, one for each question. Please staple or paperclip all pages from a single question together, but **DO NOT** staple pages from different questions together. This will make it much easier for the TA's to split up your homework for grading. Include your name and email address on each set.

1 Kernel Density Estimation [Will Bishop, 20 Points]

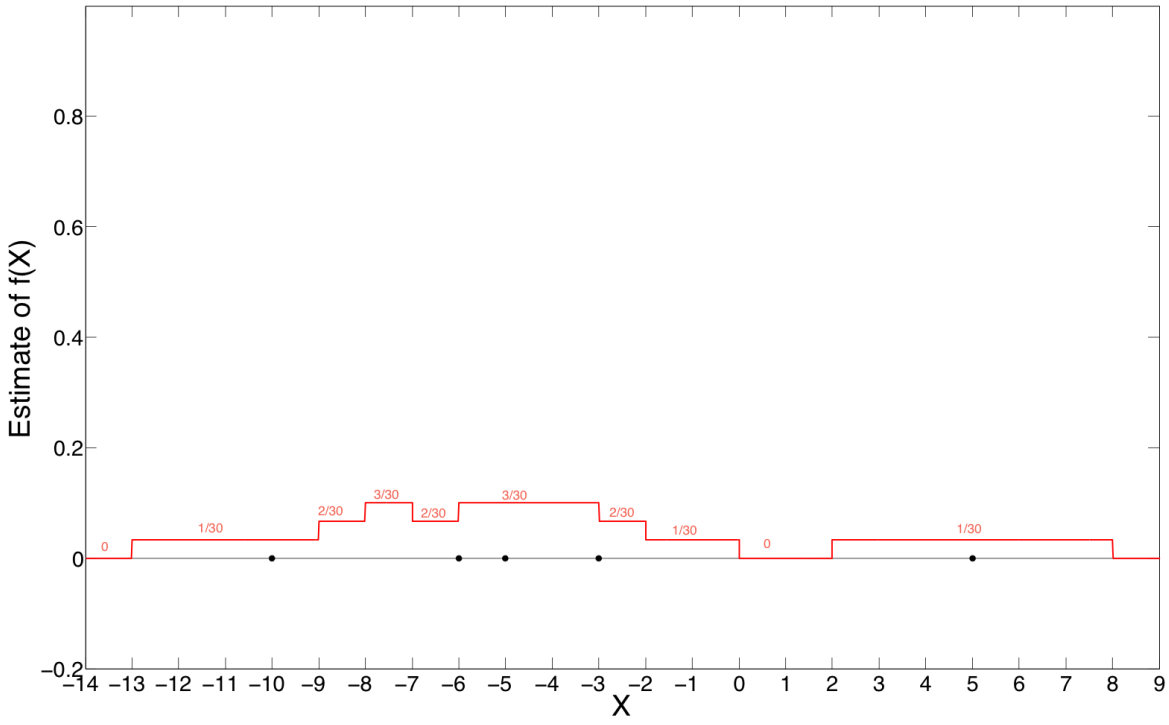
In this homework problem, we will explore kernel density estimation using a boxcar kernel. In other words, given a sample of observed data $x_1, \ldots, x_N$, we will estimate the probability density function as:

$$\hat{f}(X) = \frac{\sum_{j=1}^{N} I\left(|X - x_j| \leq h\right)}{2hN}$$

where $|\cdot|$ indicates absolute value and $I(\cdot)$ is an indicator function that takes on value 1 when the logical statement inside the parenthesis is true and 0 otherwise.

1. [3 pts] In the figure below, each dot represents a sampled data point. In this same figure, please draw $\hat{f}(x)$ for $h = 3$.

   Answer shown below.



2. [3 pts] Let $F(X)$ represent the true cumulative distribution function for a random variable X and let $f(X)$ represent the probability density function for this same variable. Please show that:

$$\mathbb{E}\left(\hat{f}(x)\right) = \frac{F(x+h) - F(x-h)}{2h}$$

Answer:

$$\begin{aligned}\mathbb{E}\left(\hat{f}(X)\right) &= \mathbb{E}\left(\frac{\sum_{j=1}^{N} I\left(|X - x_j| \leq h\right)}{2hN}\right) \\ &= \frac{1}{2hN}\mathbb{E}\left(\sum_{j=1}^{N} I\left(|X - x_j| \leq h\right)\right)\end{aligned}$$

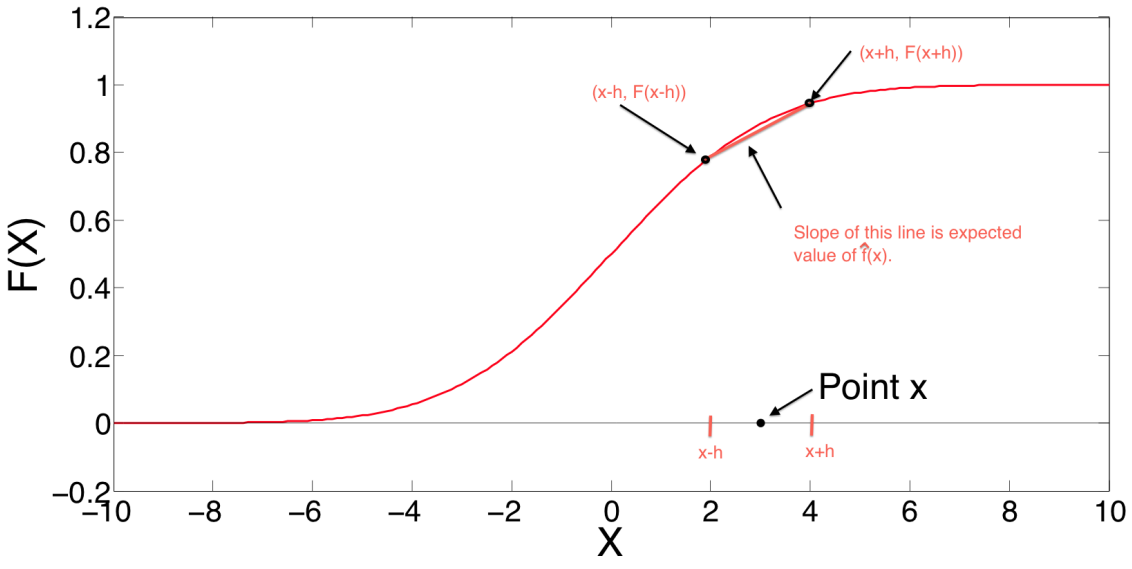
Note that the sample points are independent and identically distributed, so we can write:

$$
\begin{aligned}
\mathbb{E}\left(\hat{f}(X)\right) &= \frac{1}{2hN}\sum_{j=1}^{N}\mathbb{E}\left(I\left(|X-x_j|\leq h\right)\right) \\
&= \frac{1}{2hN}\left(N\mathbb{E}\left(I\left(|X-x|\leq h\right)\right)\right) \\
&= \frac{1}{2h}\left(\mathbb{E}\left(I\left(|X-x|\leq h\right)\right)\right) \\
&= \frac{1}{2h}P(|X-x|\leq h) \\
&= \frac{1}{2h}P(-h\leq X-x\leq h) \\
&= \frac{1}{2h}P(X-h\leq x\leq X+h) \\
&= \frac{F(x+h)-F(x-h)}{2h}
\end{aligned}
$$

3. [6 pts] The figure below shows a portion of a cumulative distribution function, $F(X)$.

   (a) Given the point x, please draw and label the points $(x+h, F(x+h))$ and $(x-h, F(x-h))$. *Answer shown below.*

   (b) Please draw a simple graphical representation relating the points $(x+h, F(x+h))$ and $(x-h, F(x-h))$ to $\mathbb{E}\left(\hat{f}(x)\right)$ and in no more than two sentences explain how this is an approximation to $f(x)$, the true value of the probability density function evaluated at x.
   Answer: We know that $f(X)$ is simply the derivative of $F(X)$. We see that $\mathbb{E}\left(\hat{f}(X)\right)$ is simply an approximation to this derivative.



4. [8 pts] Using your intuition gained to this point, please answer *true* if $\hat{f}(x)$ will be an unbiased estimator of x in the following scenarios and *false* if not. Justify your answer.

   (a) $f(X)$ is a uniform distribution between 0 and 1. $\hat{f}(.25)$ is estimated using a value of $h = .5$ *Answer: false.*

$$(F(.75) - F(-.25))/(2 \times .5) = .75 \neq 1$$

(b) $f(X)$ is a uniform distribution between 0 and 1. $\hat{f}(.25)$ is estimated using a value of $h = .2$
Answer: true

$$(F(.45) - F(.05))/(2 \times .2) = .4/.4 = 1$$

(c) $f(X) = 2X$ if $0 \leq X \leq 1$ and $f(X) = 0$ otherwise. $\hat{f}(.25)$ is estimated using a value of $h = .2$
Answer: true
The fact that the answer is true can be intuitively recognized by noting that $\mathbb{E}\left(\hat{f}(x)\right)$ can be found by calculating the slope between the points $(x-h, F(x-h))$ and $(x+h, F(x+h))$. Therefore, if the derivative of F(X) is constant in this region, the slope between these two points will in fact be the exact derivative of F(x) and there will be no bias.

(d) $f(X) = \frac{3}{2}X^2$ if $-1 \leq X \leq 1$ and $f(X) = 0$ otherwise. $\hat{f}(0)$ is estimated using a value of $h = .2$
Answer: false.
Note that f(0) = 0. Further, because f(x) takes on non-zero values between $x = -.2$ and $x = .2$, it must be that $F(.2) > F(-.2)$. Given the equation we have derived for $\mathbb{E}\left(\hat{f}(x)\right)$, it can then be immediately recognized that $\mathbb{E}\left(\hat{f}(0)\right) > 0$.

2 Support Vector Machines [Mladen Kolar, 25 points]

Suppose you are given 6 one-dimensional points: 3 with negative labels $x_1 = -1$, $x_2 = 0$, $x_3 = 1$ and 3 with positive labels $x_4 = -3$, $x_5 = -2$, $x_6 = 3$. It was shown in the class that this data cannot be separated using a linear separator. However, if the following feature map $\phi(u) = (u, u^2)$, which transforms points in $\mathbb{R}^1$ to points in $\mathbb{R}^2$, is used, a linear separator can perfectly separate the points in the new $\mathbb{R}^2$ features space induced by ϕ.

2.1 Feature Mappings

1. [2pts] Give the analytic form of the kernel that corresponds to the feature map ϕ in terms of only X_1 and X_1'. Specifically define $k(X_1, X_1')$.

   ★ **SOLUTION:**

   $$k(X_1, X_1') = X_1 X_1'(1 + X_1 X_1')$$

2. [5pt] Construct a maximum-margin separating hyperplane. This hyperplane will be a line in $\mathbb{R}^2$, which can be parameterized by its normal equation, i.e. $w_1 Y_1 + w_2 Y_2 + c = 0$ for appropriate choices of w_1, w_2, c. Here, $(Y_1, Y_2) = \phi(X_1)$ is the result of applying the feature map ϕ to the original feature X_1. Give the values for w_1, w_2, c. Also, explicitly compute the margin for your hyperplane. You do not need to solve a quadratic program to find the maximum margin hyperplane. Note that the line must pass somewhere between (-2,4) and (-1,1) (why?), and that the hyperplane must be perpendicular to the line connecting these two points. Use only two support vectors.

   ★ **SOLUTION:** To solve this problem in general, we need to solve the quadratic program. However, in this case, we will use geometric intuition. First, we observe that the line must pass somewhere between $u_1 = (-2, 4)$ and $u_2 = (-1, 1)$. This is necessary to fully separate the data and sufficient because they are the closest points of opposite class.

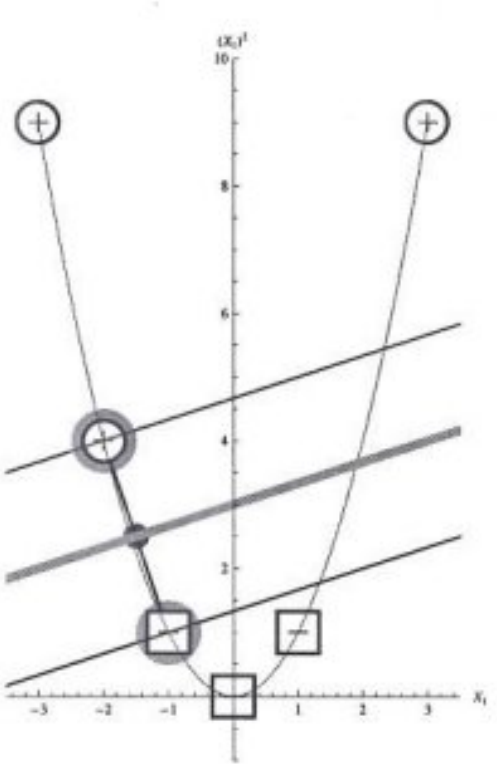
   To maximize the margin, the decision boundary needs to pass through the center $(-3/5, 5/2)$ of the line connecting u_1 and u_2. In addition it needs to be perpendicular to the line segment connecting u_1 and u_2. Now it is simple to see that the decision boundary can be described using the following equation

   $$-Y_1 + 3Y_2 - 9 = 0.$$

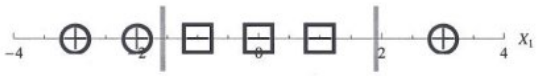
These parameters can be rescaled without affecting the equation.

The margin is simply distance between one of the canonical planes and the separating hyperplane, which is $\sqrt{10}/2$.

3. [4pt] Apply ϕ to the data and plot the points in the new $\mathbb{R}^2$ feature space. On the plot of the transformed points, plot the separating hyperplane and the margin, and circle the support vectors.



4. [2pt] Draw the decision boundary of the separating hyperplane in the original $\mathbb{R}^1$ feature space.



5. [5pt] Compute the coefficients α and the constant b in Eq. (1) for the kernel k and the support vectors $SV = \{u_1, u_2\}$ you chose in part 4. Be sure to explain how you obtained these coefficients.

$$y(x) = \text{sign}\left(\sum_{n=1}^{|SV|} \alpha_n y_n k(x, u_n) + b\right) \tag{1}$$

Think about the dual form of the quadratic program and the constraints placed on the α values.

★ **SOLUTION:** To answer this problem, we need to solve the quadratic program on the support vectors u_1 and u_2. We will use the fact that the only support vectors are u_1 and u_2, therefore only a_1 and a_2 are non-zero. Due to the constraint on the quadratic program, we know that $a_1 = a_2 = a$.

$$\begin{aligned} L(a) &= \sum_{n=1}^{N} a_n + \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m k(x_n, x_m) \\ &= a_1 + a_2 - \frac{1}{2}(a_1^2(u_1 \cdot u_1) - 2a_1 a_2 (u_1 \cdot u_2) + a_2^2 (u_2 \cdot u_2) \\ &= 2a - 5a^2. \end{aligned}$$

Taking the derivative and setting it equal to zero, we obtain $a = a_1 = a_2 = 1/5$. To compute b we recall that on the positive margin we have

$$1 = \sum_{n=1}^{|SV|} a_n y_n k(x, u_n) + b.$$

Plugging in a, we simply derive that $b = -9/5$.

6. [2pt] If we add another positive ($Y = +$) point to the training set at $X_1 = 5$ would the hyperplane or margin change? Why or why not?

★ **SOLUTION:** No. The additional point is outside the margin and is classified correctly.

2.2 Infinite Features Spaces

Lets define a new (infinitely) more complicated feature transformation $\phi_n : \mathbb{R}^1 \to \mathbb{R}^n$ given in Eq. (2).

$$\phi_n(x) = \left\{ e^{-x^2/2}, e^{-x^2/2}x, \frac{e^{-x^2/2}x^2}{\sqrt{2}}, \ldots, \frac{e^{-x^2/2}x^i}{\sqrt{i!}} \ldots, \frac{e^{-x^2/2}x^n}{\sqrt{n!}} \right\} \tag{2}$$

Suppose we let $n \to \infty$ and define new feature transformation in Eq. (3). You can think of this feature transformation as taking some finite feature vector and producing an infinite dimensional feature vector rather than the simple two dimensional feature vector used in the earlier part of this problem.

$$\phi_\infty(x) = \left\{ e^{-x^2/2}, e^{-x^2/2}x, \frac{e^{-x^2/2}x^2}{\sqrt{2}}, \ldots, \frac{e^{-x^2/2}x^i}{\sqrt{i!}} \ldots \right\} \tag{3}$$

1. [3pt] We know that we can express a linear classifier using only inner products of support vectors in the transformed feature space as seen in (1). It would be great if we could some how use the feature space obtained by the feature transformormation ϕ_∞. However, to do this we must be able to compute the inner product of examples in this infinite vector space. Lets define the inner product between two infinite vectors $a = \{a_1, \ldots, a_i, \ldots\}$ and $b = \{b_1, \ldots, b_i, \ldots\}$ as the infinite sum given in (4).

$$k(a, b) = a \cdot b = \sum_{i=1}^{\infty} a_i b_i \tag{4}$$

Can we explicity compute $k(a, b)$? What is the explicit form of $k(a, b)$? Hint you may want to use the Taylor series expansion of e^x which is given in (5).

$$e^x = \lim_{n\to\infty} \sum_{i=0}^{n} \frac{x^i}{i!} \tag{5}$$

★ **SOLUTION:**

$$
\begin{aligned}
k(a,b) &= \sum_{i=1}^{\infty} \frac{e^{-a^2/2}a^i}{\sqrt{i!}} \frac{e^{-b^2/2}b^i}{\sqrt{i!}} \\
&= \exp(-\frac{a^2+b^2}{2}) \sum_{i=1}^{\infty} \frac{(ab)^i}{i!} \\
&= \exp(-\frac{a^2+b^2}{2}) \exp(ab) \\
&= \exp(-\frac{(a-b)^2}{2}).
\end{aligned}
$$

2. [2pt] With such a high dimensional feature space should we be concerned about overfitting?

★ **SOLUTION:** Normally we would worry, however, in the case of SVMs, model complexity is managed by the number of support vectors rather than the dimensionality of the feature space.

3 Boosting [Shing-hon Lau, 25 points]

1. [4 pts] Suppose I have the 2-dimensional dataset depicted in Figure 1. Will Adaboost (with Decision Stumps as the weak classifier) ever achieve better than 50% classification accuracy on this dataset? Why or why not? Briefly justify your answer.

★ **SOLUTION:** Adaboost will never achieve better than 50% classification accuracy on this dataset since every single split produces an error of 50%.

2. [4 pts] Suppose AdaBoost is run on m training examples, and suppose on each round that the weighted training error ϵ_t of the t^{th} weak hypothesis is at most γ, for some number $0.5 > \gamma > 0$. After how many iterations, T, will the combined hypothesis H be consistent with the m training examples, i.e., achieves zero training error? Your answer should only be expressed in terms of m and γ. (Hint: What is the training error when 1 example is misclassified?)

★ **SOLUTION:** We have that the training error of the final classifier, H, is bounded by:
$\frac{1}{m}\sum_{i=1}^{m}\delta(H(x_i) \neq y_i) \leq exp(-2\sum_{t=1}^{T}(0.5-\epsilon_t)^2)$.
We can ensure that H is consistent with the m training points if we bound the training error by $\frac{1}{m}$. Thus, we would like to choose T such that:
$\frac{1}{m} \geq exp(-2\sum_{t=1}^{T}(0.5-\epsilon_t)^2)$
However, it is difficult to operate with the expression on the right-hand side, so we instead choose T such that:
$\frac{1}{m} \geq exp(-2T(0.5-\gamma_t)^2)$
Note that this is OK since $\frac{1}{m} \geq exp(-2T(0.5-\gamma_t)^2) \geq exp(-2\sum_{t=1}^{T}(0.5-\epsilon_t)^2)$
Solving gives us that:
$$\frac{1}{m} \geq exp(-2T(0.5-\gamma_t)^2)$$

$\Rightarrow ln(m) \leq 2T(0.5 - \gamma_t)^2$
$\Rightarrow \frac{ln(m)}{2(0.5-\gamma_t)^2} \leq T$

For the next six questions, consider the 1-dimensional dataset depicted in Figure 2. Suppose that we are using Adaboost with Decision Stumps as the weak classifier.

3. [3 pts] Draw the decision boundary of the first classifier, h_1. Indicate which side is classified as the + class.

   ★ **SOLUTION:** The boundary goes between the 3 '+' points and the 3 '-' points, with the '+' side on the left.

4. [3 pts] Compute and report ϵ_1 and α_1. What is the classification accuracy if we stop Adaboost here?

   ★ **SOLUTION:** $\epsilon_1 = \frac{1}{7}$ and $\alpha_1 = \frac{1}{2}ln\frac{6/7}{1/7} = \frac{1}{2}ln(6) = 0.8959$
   The classification accuracy would be $\frac{6}{7}$.

5. [3 pts] What are the new weights, $D_2(i)$, for the seven points? (Hint: Remember to normalize the weights by Z.)

   ★ **SOLUTION:** $Z_1 = \frac{6}{7}exp(-0.8959) + \frac{1}{7}exp(0.8959) = 0.6998542$
   $D_2(i) = \frac{(1/7)exp(-0.8959)}{Z_1} = 0.083333 \quad \forall i = 1, ..., 6$
   $D_2(7) = \frac{(1/7)exp(0.8959)}{Z_1} = 0.5$

6. [3 pts] Draw the decision boundary of the second classifier, h_2. Again, indicate which side is classified as the + class.

   ★ **SOLUTION:** The boundary goes between the single '+' point and the 3 '-' points, with the '+' side on the right.

7. [2 pts] Which point(s) will have the lowest weight after the second iteration of Adaboost is finished?

   ★ **SOLUTION:** The 3 '-' points will have the lowest weight after the second iteration since they were classified correctly by both h_1 and h_2.

8. [3 pts] Does the classification accuracy improve between first and second iterations of Adaboost? Explain briefly why the accuracy does (or does not) improve.

   ★ **SOLUTION:** No, the classification accuracy does not improve. In fact, all of the points are classified the same after 1 and 2 iterations.

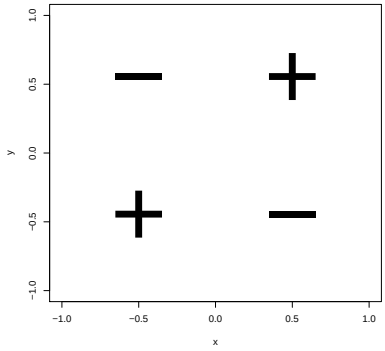


Figure 1: XOR dataset

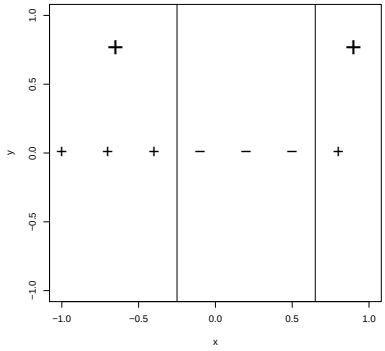


Figure 2: Seven point dataset