CS761 Spring 2017 Homework 3

Assigned Apr. 6, due Apr. 20

Instructions:

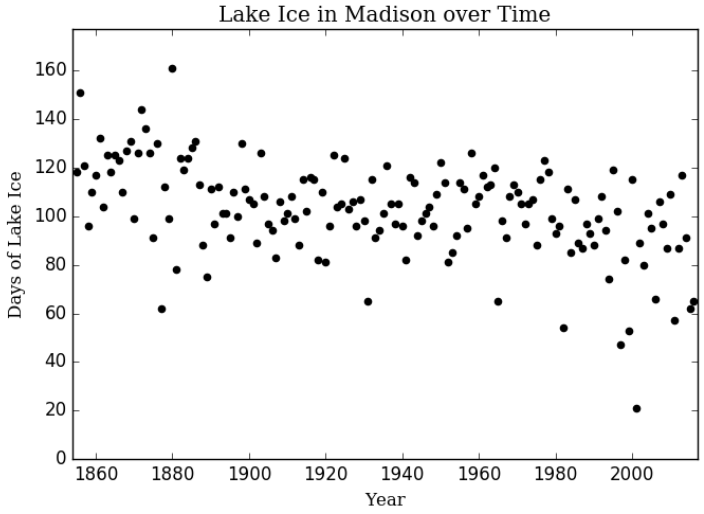
- Homeworks are to be done individually.
- Typeset your homework in latex using this file as template (e.g. use pdflatex). Show your derivations.
- Hand in the compiled pdf (not the latex file) online. Instructions will be provided. We do not accept hand-written homeworks.
- Homework will no longer be accepted once the lecture starts.
- Fill in your name and email below.

Name: David Merrell
Email: dmerrell@cs.wisc.edu

(4 questions, 25 points each)

1. The Wisconsin State Climatology Office keeps a record on the number of days Lake Mendota was covered by ice at `http://www.aos.wisc.edu/~sco/lakes/Mendota-ice.html`. The article DETERMINING THE ICE COVER ON MADISON LAKES at `http://www.aos.wisc.edu/~sco/lakes/msn-lakes_instruc.html` serves as a fine example of the Wisconsin tradition to integrate science with beer.

   (a) As with any real problems, the data is not as clean nor as organized as one would like for machine learning. Produce a clean data set starting from 1855-56 and ending in 2016-17 for the output variable DAYS. You do not need to attach your data set, but please produce a scatter plot of year vs. DAYs. Show us the sample mean and sample variance (round to 5 digits after decimal point).
   SOLUTION:



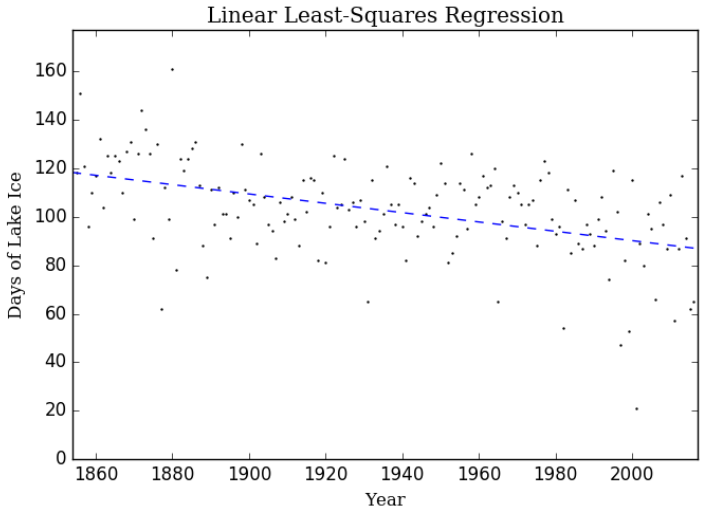
   Sample mean: **102.55556**
   (Biased) Sample Variance: **380.20988**
   (Unbiased) Sample Variance: **382.57143**

   (b) Perform ordinary least squares to estimate a linear model

$$y = \alpha + \beta x$$

   where y is DAYS and x is the year. For example, for 1855-56 the year is 1855. Show us $\hat{\alpha}, \hat{\beta}$, and an estimate of the standard error on β: $\widehat{s.e}(\hat{\beta})$.

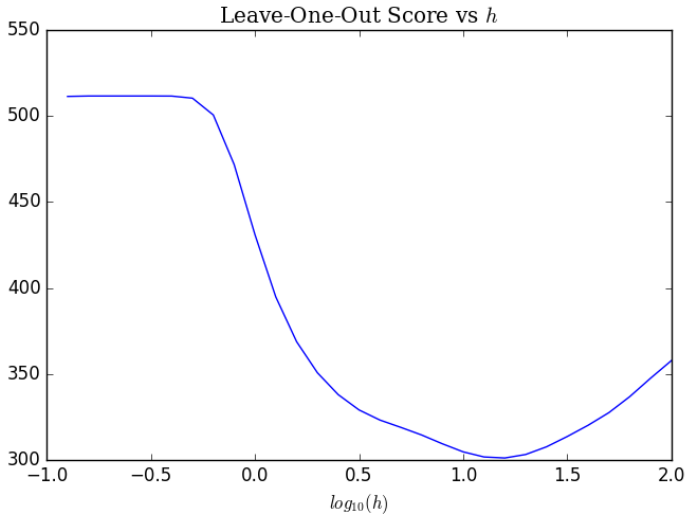
SOLUTION:



$\hat{\alpha}$: 474.99182
$\hat{\beta}$: -0.19242

(c) Perform nonparametric kernel regression using the Nadaraya-Watson estimator on this data set (input: year, output: days). Use the Gaussian kernel. Write your own code for the Nadaraya-Watson estimator. Show us the leave-one-out score (Equation 23 in lecture notes `http://pages.cs.wisc.edu/~jerryzhu/cs761/kde.pdf`) for bandwidth $h = 10^{-1}, 10^{-0.9}, 10^{-0.8}, \ldots, 10^{2}$, respectively.

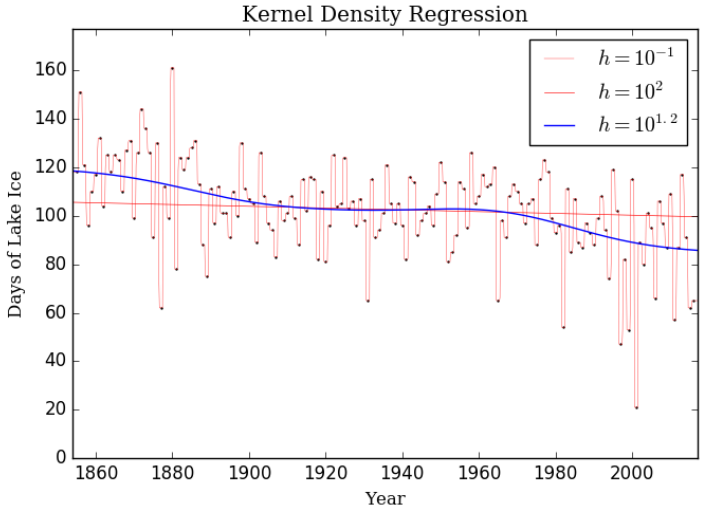
SOLUTION:



Of the bandwidths considered, $h = 1.2$ yielded the lowest leave-one-

out score.

(d) For $h = 10^{-1}, 10^2$ and the optimal h you found, respectively, plot the function estimated by Nadaraya-Watson.
SOLUTION:



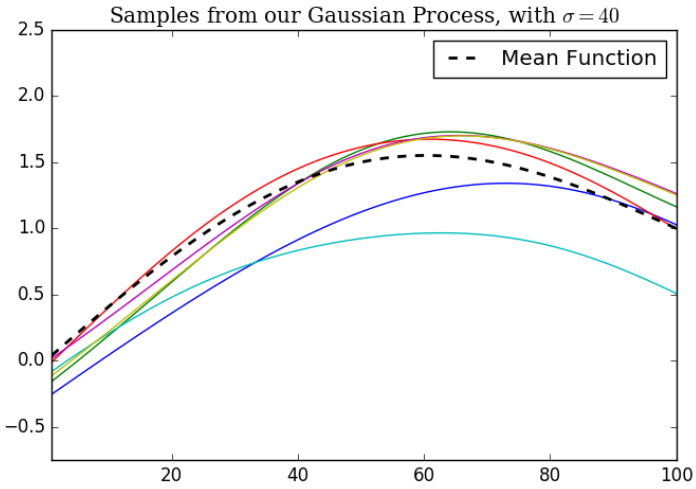
2. Consider a Gaussian Process $f \sim GP(m, k)$ over $\mathbb{R}$ with mean function

$$m(x) = \sin(\frac{\pi x}{100}) + \frac{x}{100}$$

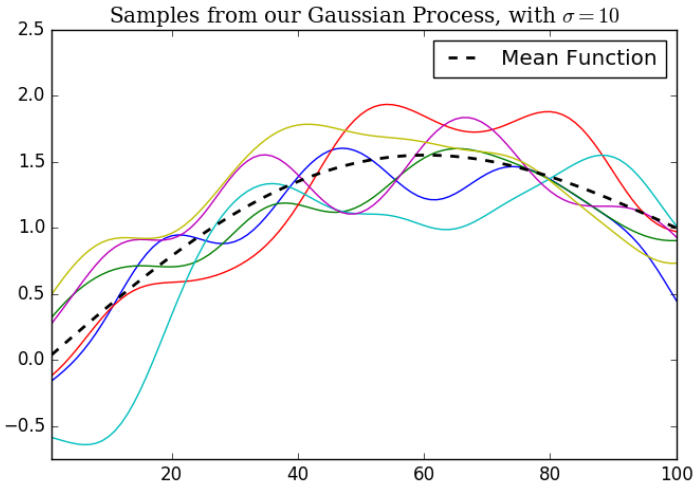
and kernel function

$$k(x, x') = \frac{1}{16} \exp\left(-\frac{(x - x')^2}{2\sigma^2}\right).$$

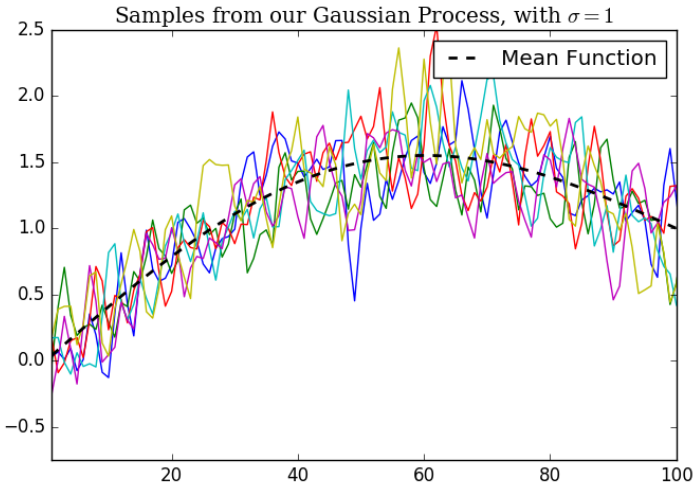
(a) Let $\sigma = 40$ (note: this is the standard deviation, not variance). Approximate the random function f by drawing $f(1), f(2), \ldots, f(100)$ from the appropriate marginal distribution. Plot the curve by connecting the dots. Show six such random functions on the same plot, together with the mean function m.
SOLUTION:



(b) Do the same with $\sigma = 10$.
SOLUTION:

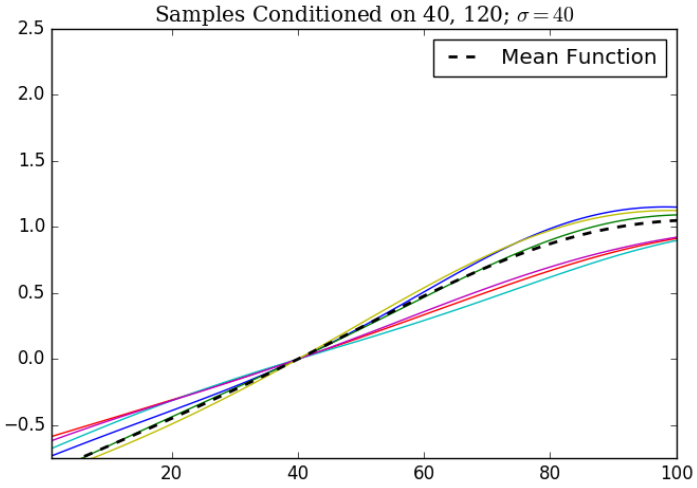


(c) Do the same with $\sigma = 1$.
SOLUTION:



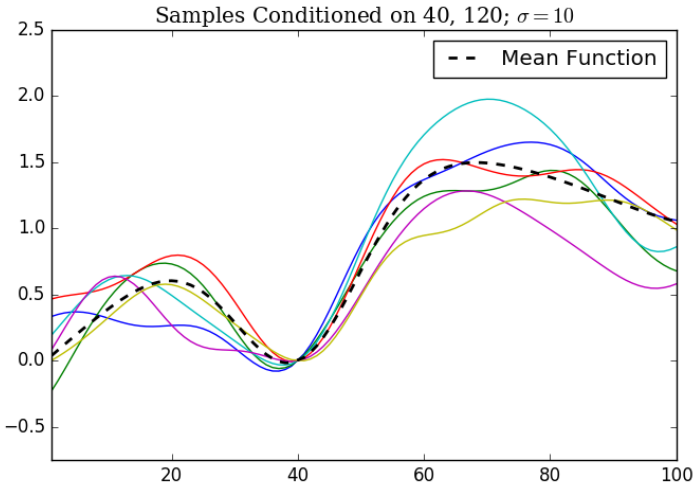
(d) Let $\sigma = 40$. Now let us observe $f(40) = 0$ and $f(120) = 1$. Now draw f from the posterior Gaussian Process conditioned on these two observations. Again, show six such f from the posterior on the same plot.

SOLUTION:

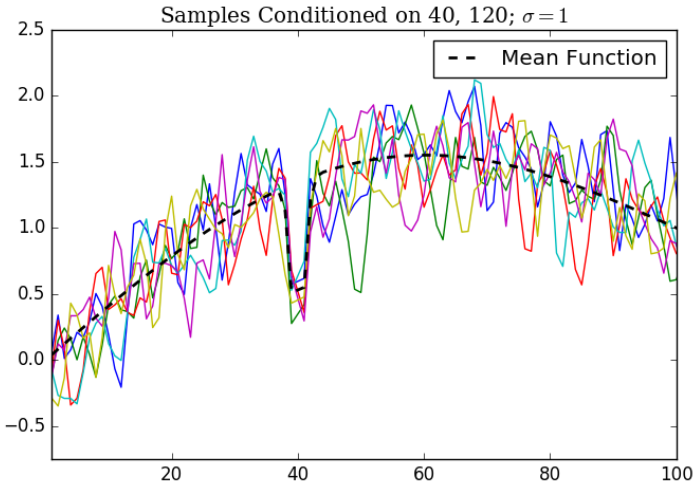


(e) Do the same with $\sigma = 10$.

SOLUTION:



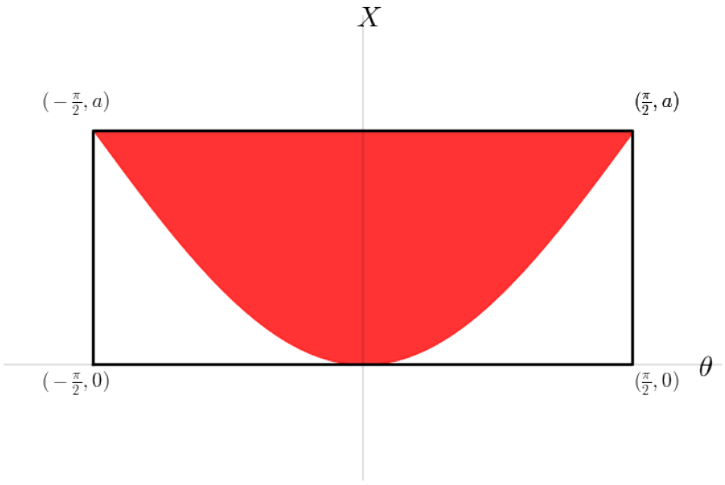
(f) Do the same with $\sigma = 1$.
SOLUTION:



3. Imagine a stick of length a. On the ground, draw parallel lines a apart. Randomly throw the stick to the ground. Each time, the stick may or may not intersect with a line.

   (a) What is the probability that the stick intersects with a line? Show your work.

SOLUTION:



We observe that the configuration of the stick with respect to the lines is fully defined by variables X and θ, where X is the displacement of one end of the stick from one of the lines, and θ is the orientation of the stick—i.e., $\theta = 0$ when the stick is perpendicular to the lines, and $\theta = \pm\frac{\pi}{2}$ when the stick is parallel to them.

If the stick is tossed randomly, then we can assume $X \sim U(0, a)$ and $\theta \sim U(-\frac{\pi}{2}, \frac{\pi}{2})$. Some trigonometry tells us that the stick crosses the *next* line when

$$X + a\cos\theta > a.$$

So the probability of the stick crossing a line is given by

$$\begin{aligned} P(\text{crossing}) &= P(X + a\cos\theta > a \\ &= P(X > a - a\cos\theta) \end{aligned}$$

Visually, this is equivalent to finding the area of the region shaded red in the above figure, and dividing it by the area of the enclosing rectangle:

$$\begin{aligned} P(X > a - a\cos\theta) &= \frac{1}{a\pi}\int_{-\pi/2}^{\pi/2} a\cos\theta d\theta \\ &= \frac{1}{\pi}\int_{-\pi/2}^{\pi/2} \cos\theta d\theta \\ &= \frac{1}{\pi}[\sin\theta]_{-\pi/2}^{\pi/2} \\ &= \frac{2}{\pi} \end{aligned}$$

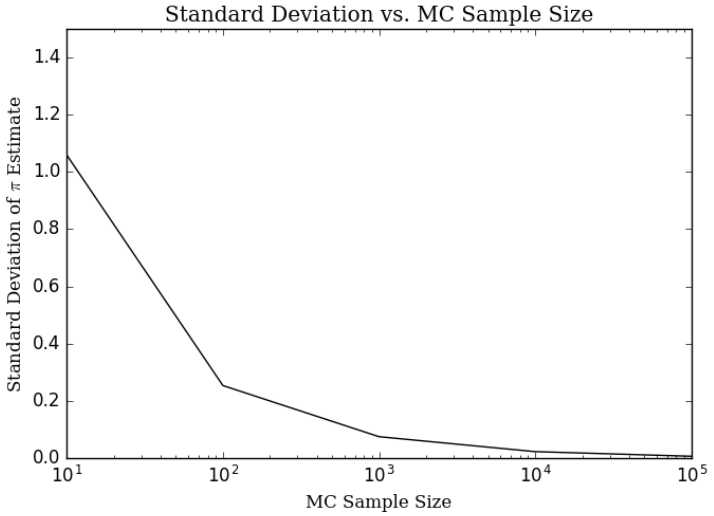
(b) Propose a Monte Carlo method for estimating π based on this.

SOLUTION: From the previous result, it follows that we can estimate π by sampling uniformly distributed X_i and θ_i, counting the fraction of instances where $X_i > a - a\cos\theta_i$, and dividing 2 by that fraction.

(c) Actually perform the experiment. Tell us about it.

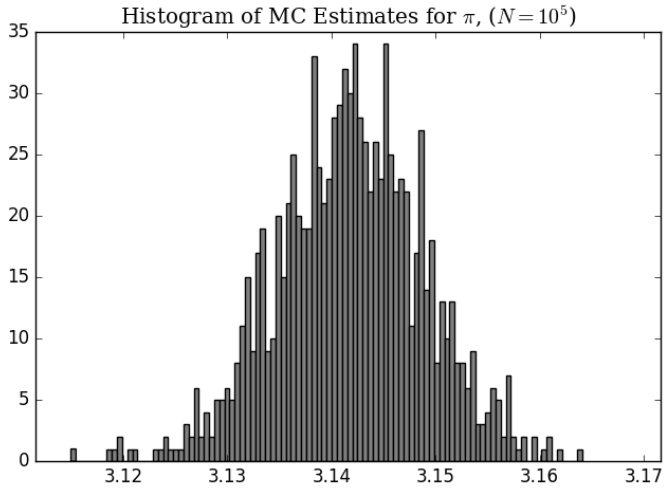
SOLUTION:

This Monte Carlo method requires very large samples in order to provide an estimate of π with any precision. For a range of 5 sample sizes, we obtained 1000 estimates for π and computed the sample standard deviation of those estimates; the result is shown in the plot below:



The main result is that standard deviation only drops to ~ 0.01 when the sample size is $\sim 10^5$; in other words, we need to throw the stick tens of thousands of times in order to reliably estimate π within two digits of precision.

We show a histogram of the estimates with a sample size of 10^5 below:



4. Consider an undirected graphical model on a binary tree with 15 nodes. Each node takes value in $\{-1, 1\}$. All edges share the same potential function $\psi(u, v) = \exp(\alpha uv)$, where u, v are a pair of parent-child nodes.

   (a) Write down the joint probability distribution defined by this graphical model.
   SOLUTION:

$$p(\vec{x}) = \frac{1}{Z(\alpha)} \prod_{(i,j)\in E} \exp(\alpha x_i x_j)$$
$$= \frac{1}{Z(\alpha)} \exp\left(\alpha \sum_{(i,j)\in E} x_i x_j\right),$$

   Where $Z(\alpha)$ is a partition function that we avoid trying to write down.

   (b) Let $\alpha = 1$. Let r be the root node and s be the left-most leaf node. Use brute force (enumerating all trees) to compute $p(r \mid s = 1)$.
   SOLUTION:
   A brute force calculation yields $p(r \mid s = 1) = 0.72087\ldots$.

   (c) Implement Gibbs sampling to estimate $p(r \mid s = 1)$. Start with the all-minus-1 tree except for $s = 1$. Go over levels in top-down order, left-to-right within each level. Discard a burn-in of 10^4 samples. Use the next 10^5 samples for estimation. Do not perform thinning.

SOLUTION:
Gibbs sampling with the prescribed burn-in and no thinning consistently under-approximates $p(r \mid s = 1)$, typically yielding numbers between 0.66 and 0.7. However, we find that thinning remedies this, yielding less-biased clustering around 0.72.

(d) Implement Metropolis-Hastings sampling to estimate $p(r \mid s = 1)$. Clearly define and discuss your proposal distribution (which has to be different than Gibbs). Use the same burn-in and number of samples as above.

SOLUTION:
We implemented Metropolis-Hastings with the following proposal distribution:

$$q(\vec{x}' \mid \vec{x}) = \frac{1}{Z} \prod_{i \in N} \exp(-x_i' x_i)$$

i.e., each node has its own independent proposal distribution,

$$q_i(x_i' \mid x_i) = \frac{1}{e + e^{-1}} \exp(-x_i' x_i).$$

In other words: q_i will propose change for x_i with probability $e/(e + e^{-1}) = .8808\ldots$. Once $\vec{x}'$ has been proposed in this way, it will be accepted with probability

$$a = \min\left(1, \frac{p(\vec{x}')}{p(\vec{x})}\right)$$

since q is symmetric.

This implementation provided estimates of $p(r \mid s = 1)$ that clustered around 0.72, even without thinning.