STAT37710/CSMSC35400 — Homework 3

University of Chicago, Spring 2024 — due Thursday, April 25 at 11:59pm

1. An online algorithm, like the perceptron, is said to be conservative if it changes its hypothesis only when it makes a mistake. Let $\mathcal{C}$ be a concept class and A be a (not necessarily conservative) online algorithm which has a finite mistake bound M on $\mathcal{C}$. Prove that there is a conservative algorithm A' for $\mathcal{C}$ which also has mistake bound M.

2. Give an example of function $k\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ that is symmetric ($k(x,x')=k(x',x)$) and positive in the sense that $k(x,x')\geq 0$ for all $x,x'\in\mathcal{X}$, but is not positive semidefinite. Conversely, give an example of a kernel that is positive semidefinite, but does not satisfy $k(x,x')\geq 0$ for all $x,x'\in\mathcal{X}$.

3. Given any function $\psi\colon \mathcal{X}\to\mathcal{X}'$, prove that if k' is a psd kernel on $\mathcal{X}'$, then $k(x,x')=k'(\psi(x),\psi(x'))$ is a psd kernel on $\mathcal{X}$.

4. Prove that if k_1 and k_2 are two positive semi-definite (psd) kernels on a space $\mathcal{X}$, then

   (a) the function $k(x,x') := k_1(x,x')+k_2(x,x')$ is a psd kernel on $\mathcal{X}$;

   (b) The function $k_{\oplus}((x_1,x_2),(x_1',x_2')) = k_1(x_1,x_1')+k_2(x_2,x_2')$ is a psd kernel on $\mathcal{X}\times\mathcal{X}$.

   (c) Given any function $\psi\colon \mathcal{X}\to\mathcal{X}'$, prove that if k' is a psd kernel on $\mathcal{X}'$, then $k(x,x')=k'(\psi(x),\psi(x'))$ is a psd kernel on $\mathcal{X}$.

   (d) Let $\alpha(\mathbf{x},\mathbf{x}')$ be the angle between $\mathbf{x},\mathbf{x}'\in\mathbb{R}^n$. Prove that the cosine kernel $k_{\angle}(\mathbf{x},\mathbf{x}')=\cos\alpha(\mathbf{x},\mathbf{x}')$ is a psd kernel on $\mathcal{X}=\mathbb{R}^n$.

5. Recall that a training set $\{(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)\}$ is said to have edge γ over a set of weak classifiers H if for any distribution D over the training set, there is at least one weak learner $h\in H$ such that $\epsilon_h=\sum_{i=1}^m D(i)\,\ell_{0/1}(h(x_i),y_i)\leq 1/2-\gamma$.

   (a) Using the inequality $\ell_{0/1}(z,1)\leq e^{-z}$ prove that after t rounds of boosting the running hypothesis $\widehat{h}(x)=\operatorname{sgn}\big(\sum_{s=1}^t \alpha_s h_s(x)\big)$ satisfies

   $$\ell_{0/1}(\widehat{h}(x_i),y_i)\leq m\big(\textstyle\prod_{s=1}^t Z_s\big)D_{t+1}(i)$$

   for every example $i=1,2,\ldots,m$.

   (b) Use this to show that

   $$\mathcal{E}_{\text{train}}(\widehat{h})\leq\prod_{s=1}^t 2\sqrt{\epsilon_s(1-\epsilon_s)}$$

   .

   (c) By plugging into the definition of the edge at round s, which is $\gamma_s=1/2-\epsilon_s$ and using the inequality $1-z\leq e^{-z}$ prove that the training error decreases exponentially,

   $$\mathcal{E}_{\text{train}}(\widehat{h})\leq\exp\left(-2\gamma^2 t\right),$$

   as stated in a Theorem in class.

6. Recall that a Gaussian Process is a distribution over functions, not just over some finite collection of variables. Specifically, a GP $\mathcal{G}(\mu,k)$ on the real line is that distribution for which if we fix n points $x_1,x_2,\ldots,x_n\in\mathbb{R}$ and draw a function f from $\mathcal{G}(\mu,k)$, the function values $f(x_1),f(x_2),\ldots,f(x_n)$ are jointly normally distributed with

   $$\begin{aligned}\mathbb{E}(f(x_i)) &= \mu(x_i),\\ \operatorname{Cov}(f(x_i),f(x_j)) &= k(x_i,x_j).\end{aligned}$$

   Here, μ and k are considered parameters of the GP, just like the vector mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ are parameters of the normal distribution: μ can be any function $\mu\colon\mathbb{R}\to\mathbb{R}$, and

$k\colon \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ can be any positive semi-definite kernel on $\mathbb{R}$. (It takes a little bit of theoretical work to show that for any μ and k, $\mathcal{G}(\mu,k)$ really is a valid distribution over functions, and that it is essentially unique.) For simplicity, in the following we will take $\mu(x)=0$, and set k to be the our favorite kernel, the Gaussian RBF kernel $k(x,x')=e^{-(x-x')^2/(2\tau^2)}$.

In the Bayesian framework, GPs are used as a prior for the function $\widehat{f}$ that we are trying to estimate. The beauty of GPs is that there are several complicated looking things that one can do with them with in a very simple way:

(i) We can draw functions from the prior $\mathcal{G}(\mu,k)$.

(ii) In a regression setting, if we assume that the observed data $S=\{(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)\}$ are distributed around f according to a second univariate Gaussian:

$$y=f(x)+\eta \qquad \eta\sim\mathcal{N}(0,\sigma^2), \tag{1}$$

we can invoke Bayes' rule

$$p(f|S)=\frac{p(S|f)\,p(f)}{p(S)}$$

to get the posterior distribution over f. Miraculously, in this case the posterior also turns out to be a Gaussian Process, $\mathcal{G}(\mu',k')$.

(iii) We can draw further functions from this updated GP $\mathcal{G}(\mu',k')$, or just use its mean $\widehat{f}=\mu'$ as our regression estimate (which will be the same as doing Ridge Regression), and k' as a measure of uncertainty about $\widehat{f}$.

In this problem you are asked to do the following:

(a) Assume that for $x_1,x_2,\ldots,x_n\in\mathbb{R}$, the function values $f(x_1),\ldots,f(x_n)$ are known, and we want to estimate the value of f at some additional point x. Define the Gram matrix $\boldsymbol{K}\in\mathbb{R}^{n\times n}$, the vector $\boldsymbol{k}_x\in\mathbb{R}^n$, and the scalar κ_x as

$$\begin{aligned}K_{i,j}&=k(x_i,x_j)\\ [k_x]_i&=k(x,x_i)\\ \kappa_x&=k(x,x),\end{aligned}$$

and let $\boldsymbol{f}=(f(x_1),\ldots,f(x_n))^\top$. Assume for simplicity that $\mu(x)=0$. The key to using GP's is that since $(f(x_1),\ldots,f(x_n),f(x))$ are jointly normal with covariance matrix

$$\boldsymbol{K}^{(n+1)}=\left[\begin{array}{c|c}\boldsymbol{K} & \boldsymbol{k}_x\\ \hline \boldsymbol{k}_x & \kappa_x\end{array}\right],$$

the distribution of $f(x)$ given $(f(x_1),\ldots,f(x_n))$ is also normal. Show that the mean and variance of $f(x)$ given $(f(x_1),\ldots,f(x_n))$ is

$$\begin{aligned}\mathbb{E}(f(x))&=\boldsymbol{k}_x^\top\boldsymbol{K}^{-1}\boldsymbol{f}\\ \mathrm{Var}(f(x))&=\kappa_x-\boldsymbol{k}_x^\top\boldsymbol{K}^{-1}\boldsymbol{k}_x.\end{aligned}$$

(b) Similarly, show that if $y_1,\ldots,y_n$ are distributed around $x_1,\ldots,x_n$ according to (**??**), then given $\boldsymbol{y}=(y_1,y_2,\ldots,y_n)$,

$$p(f(x)|y_1,\ldots,y_n)\sim\mathcal{N}\big(\boldsymbol{k}_x^\top(\boldsymbol{K}+\sigma^2\boldsymbol{I})^{-1}\boldsymbol{y}\,,\;\kappa_x-\boldsymbol{k}_x^\top(\boldsymbol{K}+\sigma^2\boldsymbol{I})^{-1}\boldsymbol{k}_x\big). \tag{2}$$

(c) Let $\mu(x)=0$ and $k(x,x')=e^{-(x-x')^2/(2\tau^2)}$ with $\tau^2=0.12$. Draw 20 different samples from $\mathcal{G}(\mu,k)$ and plot them, restricted to the unit interval $[0,1]$ on the x axis. For this, all that you need to do is let $z_1,\ldots,z_N$ be a sufficient number of equispaced points on $[0,1]$, and plot the line connecting $f(z_1),\ldots,f(z_N)$, where $f\sim\mathcal{G}(\mu,k)$.

(d) Now apply GP regression to the dataset `gp.dat`. Your prior should be $\mathcal{G}(\mu, k)$, as before. Given the data in `gp.dat` (the first column are the x values and the second column are the y values), plot the posterior mean

$$\mu'(x) = \boldsymbol{k}_x^\top (\boldsymbol{K} + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{y}$$

and the two standard deviation bounds around it

$$s'_\pm(x) = \boldsymbol{k}_x^\top (\boldsymbol{K} + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{y} \pm 2\sqrt{\kappa_x - \boldsymbol{k}_x^\top (\boldsymbol{K} + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{k}_x}.$$

(e) Plot 20 samples from the posterior GP $\mathcal{G}(\mu', k')$. Similarly to part (c).

Along with your results, please submit your code for each of the above parts.