ES 201 Homework 4, Due 03/28/2017 (in class)

Reading: Secs. 11.1, 11.5, 11.6, 11.7, 11.9, 11.10, 12.1, 12.2, 12.5, 12.6, 12.7, Lecs. 12, 13.

Kernels and kernel regression

Problem 1

(Variation on Problem 11.19)
Let $(s_i)_{i=1}^n$ be the discrete-time signal that represents an audio file obtained by sampling a continuous-time waveform $s(t)$ with a sampling frequency f_s. Let $(y_i)_{i=1}^n$ be a noisy version of this signal

$$y_i = s_i + \epsilon_i, \tag{1}$$

observed in additive i.i.d. Gaussian noise $(\epsilon_i)_{i=1}^n \sim \mathcal{N}(0, \sigma_\epsilon^2)$. Define the signal-to-noise ratio (SNR) in dB as

$$\text{SNR} = 10 \times \log_{10}\left(\frac{\sigma_s^2}{\sigma_\epsilon^2}\right), \tag{2}$$

and $\sigma_s^2 = \frac{1}{n}\sum_{i=1}^n (s_i - \bar{s})^2$, $\bar{s} = \frac{1}{n}\sum_{i=1}^n s_i$.

In this problem, we will de-noise a noisy version of the audio recording using kernel ridge regression. In particular, we consider the following model

$$y_i = g(x_i) + \epsilon_i,\ i = 1, \cdots, n. \tag{3}$$

where, for $i = 1, \cdots, n$, $x_i = \frac{i}{f_s}$ in seconds, and $g(\cdot)$ is a function in the Reproducing Kernel Hilbert space induced by the Gaussian kernel

$$\kappa(x_i, x_j) = \exp\left(\frac{(x_i - x_j)^2}{2\sigma^2}\right) \tag{4}$$

Read an audio file–two files, `Ed_Murrow.wav` and `Blade_Runner.wav` are a provided–using

```
import scipy.io.wavfile as wav
(fs,src) = wav.read('filename.wav')
```

Take 100 data samples (spaced 10 samples apart) from one of the channels in the file, starting from the $100,000^{\text{th}}$ sample. Then add i.i.d. Gaussian noise at a 15 dB level and "hit" 10 of the data samples with outliers (set the outlier values to 80% of the maximum value of the data samples).

(a) Compute the reconstructed data samples from the noisy signal using kernel ridge regression. Use the Gaussian kernel with $\sigma = 0.004$, and set the regularization parameter (C in book, $\lambda/2$ in lecture notes) to 0.0001.

(b) Repeat part (a) with regularization parameter values of $10^{-5}, 0.001, 0.05$.

(c) Repeat part (a) with $\sigma = 0.001, 0.01, 0.05$.

(d) Comment on your results.

Bayes and EM

Problem 2

A random variable τ is said to be distributed according to the Inverse-Gamma distribution if

$$p(\tau) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\frac{\beta}{\tau}}, \tag{5}$$

for constants $\alpha > 0$ and $\beta > 0$.

Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$ with μ known. Let $\tau = \frac{1}{\sigma^2}$. Show that the Inverse-Gamma distribution is a conjugate prior for the likelihood from a Gaussian distribution with respect to the precision τ.

Problem 3

Problem 12.3.

Problem 4

The Old Faithful geyser of Yellowstone National Park is one of the most famous geysers in the world. A geyser can be described simply as a hot spring in which water intermittently boils, thus sending a tall column of water and steam into the air, in a spectacular fashion! For the purposes of this exercise, the following is important to understand regarding the basic Physics of geysers such as old faithful.

1. The internals of a geyser can be thought of as a chamber, that builds up water, which is then heated up by magma beneath, followed by the water being expelled after enough pressure has built up inside the chamber.

2. Not all of the water in the chamber is expelled during an eruption.

3. **The less water is expelled during an eruption, the less time it takes for the chamber to replenish itself, and therefore the less time until the next eruption occurs.**

Therefore, we expect the Old Faithful geyser behavior to exhibit at least two clusters in the eruption-duration vs time-to-next-eruption space, and can use this to build a predictor for the time until the next eruption, given the duration of the previous eruption.

Consider the set $\{\mathbf{x}_i = ((x_{i,1}, x_{i,2})\}_{i=1}^n\}$ of all Old Faithul eruptions in a given year, where $x_{i,1}$ is the duration of an eruption and $x_{i,2}$ the time until the *next* eruption. Figure 1 is a plot of $x_{1,i}$ against $x_{2,i}$ for two different years. The figure suggest that, if we can split the data into two sets–one for which the eruption duration is "short" and another for which it is "long"–the time until the next eruption can be linearly predicted from the duration of the previous eruption. For each year, we can treat the data in

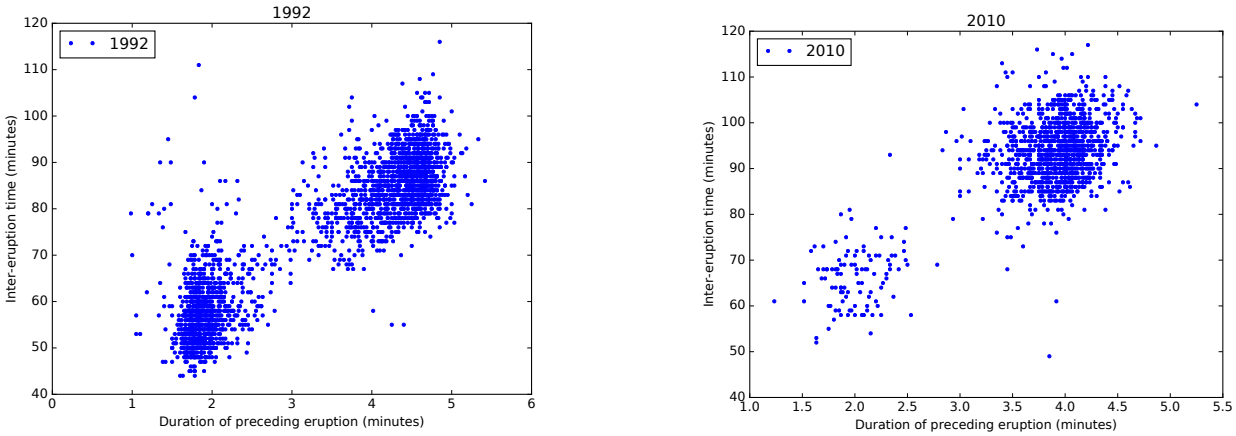

Figure 1: Relationship between duration of an eruption and time until the following eruption for the Old Faithful geyser.

Figure 1 as coming from a *mixture of linear regression models*. Suppose, $\theta \in \{0, 1\}$ is an indicator for whether the previous eruption was "short" ($\theta = 0$) or "long" ($\theta = 1$) and $P[\theta = 1] = \pi$. Our *mixture of linear regression models (MLRMs)* is specified as under a mixture of Gaussian model

$$\theta_i \sim \text{Bernoulli}(\pi) \tag{6}$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2_{\theta_i}) \tag{7}$$

$$x_{2,i} = \alpha_{\theta_i} x_{1,i} + \beta_{\theta_i} + \epsilon_i, \tag{8}$$

where $\zeta = (\pi, \alpha_0, \beta_0, \sigma_0^2, \alpha_1, \beta_1, \sigma_1^2)$ are a set of unknown parameters we would like to estimate from eruption data. Given θ_i, the model asserts that the time until the next (i^{th}) eruption can be linearly predicted using the duration of the previous eruption, hence the name MLRMs. For a given year, we treat the data $\{(x_{2,i}, \theta_i)\}_{i=1}^n$ (the *complete data* in EM terminology) as i.i.d. samples from the MLRMs. We treat $x_{1,i}$ as a covariate that can be used to predict when the next eruption will occur.

N.B.: We are assuming that $\{x_{1,i}\}_{i=1}^n$ are given, so that the conditioning on their values is implicit in all probability densities to follow.

(a) Show that the *marginal likelihood* of $x_{2,i}$ given $x_{1,i}$ is

$$p(x_{2,i}; \zeta) = \pi \cdot \mathcal{N}(x_{2,i}; \alpha_1 x_{1,i} + \beta_1, \sigma_1^2) + (1 - \pi) \cdot \mathcal{N}(x_{2,i}; \alpha_0 x_{1,i} + \beta_0, \sigma_0^2). \tag{9}$$

(b) Show that, given $x_{1,i}$, the *complete data likelihood* of the pair $(x_{2,i}, \theta_i)$ is

$$p(x_{2,i}, \theta_i; \zeta) = \big(\pi \cdot \mathcal{N}(x_{2,i}; \alpha_1 x_{1,i} + \beta_1, \sigma_1^2)\big)^{\theta_i} \cdot \big((1 - \pi) \cdot \mathcal{N}(x_{2,i}; \alpha_0 x_{1,i} + \beta_0, \sigma_0^2)\big)^{1-\theta_i}. \tag{10}$$

(c) Suppose $\{x_{1,i}\}_{i=1}^n$ are given, and that $\{(x_{2,i}, \theta_i)\}_{i=1}^n$ are i.i.d. samples from the MLRMs, give an expression for $\log p(\{(x_{2,i}, \theta_i)\}_{i=1}^n; \zeta)$ as a function of ζ.

(d) We will use EM to maximize $\log p(\{(x_{2,i})\}_{i=1}^n; \zeta)$. Suppose $\zeta^{(\ell)}$ is a guess of the maximum likelihood estimate of ζ. Give an expression for $\mathcal{Q}(\zeta|\zeta^{(\ell)})$ and show that the *sufficient statistic* of the E-step is $P[\theta_i|x_{2,i}, \zeta^{(\ell)}]$. Give an explicit expression for $P[\theta_i|x_{2,i}, \zeta^{(\ell)}]$ in terms of $\zeta^{(\ell)}$.

Hint: Recall that we are assuming that $\{x_{1,i}\}_{i=1}^n$ are given, so that the conditioning on their values is implicit in $P[\theta_i|x_{2,i}, \zeta^{(\ell)}]$.

We will now proceed to the M-step of EM, where we derive expressions for $\zeta^{(\ell+1)}$.

(e) Show that $(\alpha_0^{(\ell+1)}, \beta_0^{(\ell+1)})$ and $(\alpha_1^{(\ell+1)}, \beta_1^{(\ell+1)})$ each are the solutions of a weighted least-squares problem. Derive the update equation for $\pi^{(\ell+1)}$. Derive expressions for $\sigma_0^{2(\ell+1)}$ and $\sigma_1^{2(\ell+1)}$ *assuming* $(\alpha_0^{(\ell+1)}, \beta_0^{(\ell+1)})$ and $(\alpha_1^{(\ell+1)}, \beta_1^{(\ell+1)})$ are known respectively.

(f) Apply the EM algorithm you have derived to the eruption data from Old Faithful (`PrevErDur1992.json`, `Time2NextEr1992.json`) and (`PrevErDur2010.json`, `Time2NextEr2010.json`). For each year, the files are dictionaries, indexed by date, whose values are lists of eruption durations and inter-eruption times respectively. Reproduce the plots of Figure 1 and, for each class, plot the regression line for the class. You may find the file `GeyserDataRead.py` useful.

Hint: Explore various forms of initializations for your EM algorithm. To debug your algorithm, first "cheat" and visually assign the data to one of the classes by applying a threshold to the x_1 axis. Second, try initializing your algorithm randomly.

(g) EM guarantees that log likelihood $\log p(\{x_{2,i}\}_{i=1}^n; \zeta^{(\ell)})$ should be non-decreasing. Plot the log likelihood as a function of ℓ.

Problem 5 (Extra credit)

Problem 12.12.