Quiz 1

15 Marks, 45 minutes

Thursday 25$^{\text{th}}$ August, 2016

Please answer **to the point** in the limited space provided for each question. You can do rough work in a separate sheet of paper provided to you. You can also assume any result stated or proved in the class (but NOT as part of the tutorials).

Problem 1. Let $\mathcal{D} = \langle(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\rangle$ such that each $y_j \in \Re$. Let $\phi(\mathbf{x}) = [\phi_1(\mathbf{x}), \ldots, \phi_n(\mathbf{x})]$ be a vector of basis functions. Consider the linear regression function $f(\mathbf{x}) = \phi^T(\mathbf{x})\mathbf{w}$ with $\mathbf{w}$ obtained either as a least squares or ridge regression estimate. Show that, using either of these estimates for $\mathbf{w}$, the regression function can be written in the (so-called *kernelized*) form $f(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i K(\mathbf{x}, \mathbf{x}_i) y_i$ where $K(\mathbf{x}, \mathbf{x}_i) = \phi^T(\mathbf{x})\phi(\mathbf{x}_i)$ is a function of $\mathbf{x}$ and $\mathbf{x}_i$ only and independent of any of the $\mathbf{y}_i$'s and $\mathbf{x}_j$ for all $j \neq i$. Each α_i can be a function of the entire dataset $\mathcal{D}$.

Hint: Use the following Matrix Identity that holds for any matrices P, B and R with compatible dimensions such that R and $BPB^T + R$ are invertible:

$$(P^{-1} + B^T R^{-1} B)^{-1} B^T R^{-1} = PB^T (BPB^T + R)^{-1}$$

(**8 Marks**)

Answer: The solution to linear (set $\lambda = 0$) and ridge regression can be written as $\mathbf{w} = (\Phi^T\Phi + \lambda I)^{-1}\Phi^T\mathbf{y}$ where

- Recall for Ridge Regression: $\mathbf{w} = (\Phi^T\Phi + \lambda I)^{-1}\Phi^T y$, where,

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \ldots \\ y_m \end{bmatrix}$$

$$\Phi = \begin{bmatrix} \phi_1(\mathbf{x}_1) & \ldots & \phi_p(\mathbf{x}_1) \\ \ldots & \ldots & \ldots \\ \phi_1(\mathbf{x}_m) & \ldots & \phi_p(\mathbf{x}_m) \end{bmatrix}$$

- **Please note the difference between Φ and $\phi(\mathbf{x})$**

$$\phi(\mathbf{x}_j) = \begin{bmatrix} \phi_1(\mathbf{x}_j) \\ \ldots \\ \phi_p(\mathbf{x}_j) \end{bmatrix}$$

Then, the regression function will be

$$f(\mathbf{x}) = \phi^T(\mathbf{x})\mathbf{w} = \phi^T(\mathbf{x})(\Phi^T\Phi + \lambda I)^{-1}\Phi^T\mathbf{y}$$

- $\phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j) = K(\mathbf{x}_i, \mathbf{x}_j)$
- $\left(\Phi^T\Phi\right)_{ij} = \sum_{k=1}^{m} \phi_i(\mathbf{x}_k)\phi_j(\mathbf{x}_k)$
- $\left(\Phi\Phi^T\right)_{ij} = \sum_{k=1}^{p} \phi_k(\mathbf{x}_i)\phi_k(\mathbf{x}_j) = \phi^T(x_i)\phi(x_j) = K(\mathbf{x}_i, \mathbf{x}_j)$

Kernelizing Ridge Regression

- Given $\mathbf{w} = (\Phi^T\Phi + \lambda I)^{-1}\Phi^T\mathbf{y}$ and using the identity $(P^{-1} + B^T R^{-1} B)^{-1} B^T R^{-1} = PB^T(BPB^T + R)^{-1}$
  - $\Rightarrow$ by setting $R = I$, $P = \frac{1}{\lambda}I$ and $B = \Phi$,
  - $\Rightarrow$ $\mathbf{w} = \Phi^T(\Phi\Phi^T + \lambda I)^{-1}\mathbf{y} = \sum_{i=1}^{m} \alpha_i \phi(\mathbf{x}_i)$ where $\alpha_i = \left((\Phi\Phi^T + \lambda I)^{-1}\mathbf{y}\right)_i$
  - $\Rightarrow$ the final decision function $f(\mathbf{x}) = \phi^T(\mathbf{x})\mathbf{w} = \sum_{i=1}^{m} \alpha_i \phi^T(\mathbf{x})\phi(\mathbf{x}_i)$

The Kernel function in Ridge Regression

- We call $\phi^\top(x_1)\phi(x_2)$ a **kernel function**:
  $K(x_1, x_2) = \phi^\top(x_1)\phi(x_2)$
- The preceding expression for decision function becomes $f(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i K(\mathbf{x}, \mathbf{x}_i)$
  where $\alpha_i = (([K(\mathbf{x}_i, \mathbf{x}_j)] + \lambda I)^{-1}\mathbf{y})_i$

Problem 2. Case for non-IID dataset:

In the class, we discussed the case of Bayesian estimation for a univariate Gaussian from dataset $\mathcal{D}$ that consisted of IID (independent and identically distributed) observations.

Let $\Pr(X) \sim \mathcal{N}(\mu, \sigma^2)$ and let σ^2 be known. Suppose, the examples $x_1...x_m$ in the dataset $\mathcal{D}$ were not necessarily independent and whose possible dependence was expressed by known covariance matrix Ω but with a common unknown (to be estimated) mean $\mu \in \Re$. Let $\mathbf{u} = [1, 1, \dots 1]$ a $m-$dimensional vector of 1's and $\mathbf{x} = [x_1...x_m]$ and

$$Pr(x_1...x_m; \mu, \Omega) = \frac{1}{(2\pi)^{\frac{m}{2}} |\Omega|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu\mathbf{u})^T \Omega^{-1}(\mathbf{x}-\mu\mathbf{u})}$$

Assume that $\Omega \in \Re^{m \times m}$ is positive-definite. Now answer the following questions

1. How would you go about doing Bayesian estimation for μ? What will be an appropriate conjugate prior? What will the posterior be? And what will be the MAP and Bayes estimates?

2. Is the case of IID data set $\mathcal{D}$ a special case of this problem? Prove your claim.

   (**7 Marks**)

   Answer:

   This problem is directly adapted from Tutorial 2.

   Answers to 1: As hinted in the class, we will expect the conjugate prior of mean μ of the (product of) Gaussian to be Gaussian. Let $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$ with a fixed and known σ_0^2.

   $\mathcal{N}(\mu_m, \sigma_m^2) = exp\left(\frac{-1}{2\sigma_m^2}(\mu - \mu_m)^2\right) = \Pr(\mu|\mathcal{D}) \propto \Pr(\mathcal{D}|\mu)\Pr(\mu) =$

   $\frac{1}{(2\pi)^{\frac{m}{2}}|\Omega|^{\frac{1}{2}}} \frac{1}{\sqrt{2\pi\sigma_0^2}} exp\left(-\frac{1}{2}(\mathbf{x} - \mu\mathbf{u})^T\Omega^{-1}(\mathbf{x} - \mu\mathbf{u}) - \frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right) \propto exp\left(-\frac{1}{2}(\mathbf{x} - \mu\mathbf{u})^T\Omega^{-1}(\mathbf{x} - \mu\mathbf{u}) - \right.$

   Our reference equality is:

   $$exp\left(-\frac{1}{2}(\mathbf{x}^T\Omega^{-1}\mathbf{x} - 2\mu\mathbf{x}^T\Omega^{-1}\mathbf{u} + \mu^2\mathbf{u}^T\Omega^{-1}\mathbf{u}) - \frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right) = exp\left(\frac{-1}{2\sigma_m^2}(\mu - \mu_m)^2\right)$$

   Matching coefficients of μ^2, we get

   $$\frac{-\mu^2}{2\sigma_m^2} = \frac{-1}{2}\mu^2\mathbf{u}^T\Omega^{-1}\mathbf{u} + \frac{-\mu^2}{2\sigma_0^2} \Rightarrow \frac{1}{\sigma_m^2} = \frac{1}{\sigma_0^2} + \mathbf{u}^T\Omega^{-1}\mathbf{u}$$

   Matching coefficients of μ, we get

   $$\frac{2\mu\mu_m}{2\sigma_m^2} = \mu\left(\mathbf{x}^T\Omega^{-1}\mathbf{u} + \frac{2\mu_0}{2\sigma_0^2}\right) \Rightarrow \mu_m = \sigma_m^2\left(\mathbf{x}^T\Omega^{-1}\mathbf{u} + \frac{\mu_0}{\sigma_0^2}\right) \Rightarrow \frac{1}{1+\sigma_0^2\mathbf{u}^T\Omega^{-1}\mathbf{u}}\left(\sigma_0^2\mathbf{x}^T\Omega^{-1}\mathbf{u} + \mu_0\right)$$

   μ_m will be the MAP estimate of μ.

   One can easily verify that setting $\Omega = I$ gives the IID case. In fact, one can also verify that the with $\Omega = I$, one gets the same Bayesian estimates as in the IID case discussed in the class.