# ECE 592 – Topics in Data Science

Test 4: Machine Learning – Fall 2020

October 28, 2020

Please remember to justify your answers carefully.

Last name: _____ First name: _____

Please recall the course academic integrity policy for tests:
No cooperation or "collaboration" between students is allowed. Especially during an online course experience, it could be tempting to text or email a friend. This is not allowed. You will be allowed to use your notes, books, a browser, and software such as Matlab and/or Python.[1] However, while working on the test you should not text, email, or communicate with other people (certainly not other students) in any way, unless you are consulting with the course staff. **By submitting the test, you will be acknowledging that you completed the work on your own without the help of others in any capacity.** Any such aid would be unauthorized and a violation of the academic integrity policy.

[1] You can use the browser to access Moodle, the course webpage, and look up technical topics. Similar to a normal test, you must not communicate with other people.

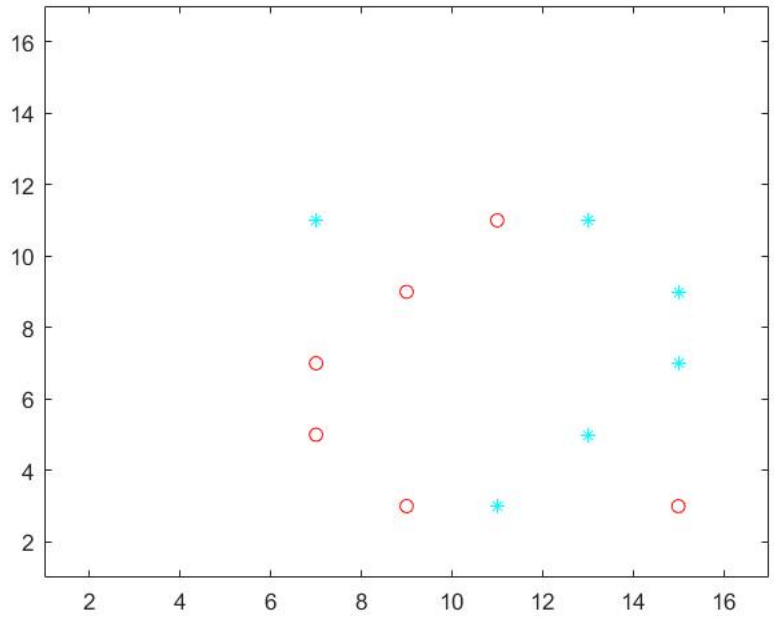
**Question 1** (Bayesian classification)
Consider a Bayesian classification problem where there are two classes, red and blue. The probabilities of the two classes are $Pr(\text{blue}) = 0.6$ and $Pr(\text{red}) = 0.4$. A random variable (RV) $X$ is generated in different ways based on the class. For each class, the conditional probability density function (pdf) is a Gaussian mixture with 2 components, $f_{\text{blue}} = 0.5\mathcal{N}(0,1) + 0.5\mathcal{N}(1,1)$ and $f_{\text{red}} = 0.8\mathcal{N}(0,1) + 0.2\mathcal{N}(1,1)$, where the means of all Gaussian components are 0 and 1, the variances corresponding to all Gaussian components are 1, and we have probabilities 0.2, 0.5, and 0.8 for the components. Recall that a Gaussian RV $X$ with mean $\mu$ and variance $\sigma^2$ has a pdf given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

**(a)** Derive a Bayesian classifier by computing the posterior probability, $Pr(\text{red}|x)$.

**(b)** Find the decision boundary/boundaries for the classifier that you derived above? (There is no need to provide numerical values; an expression is fine.)

**Question 2** (Nearest neighbors classification)
You are given training vectors labeled into two classes, where each vector is in $p = 2$ dimensions. The training set is comprised of the following points, which are plotted below.
Class 1 (red circles): $[(11, 11); (15, 3); (9, 9); (7, 7); (7, 5); (9, 3)]$
Class 2 (blue asterisks): $[(15, 9); (15, 7); (13, 5); (13, 11); (11, 3); (7, 11)]$




Please answer the following questions.
**(a)** For the $K$ nearest neighbors classifier, what value of $K$ seems reasonable for this dataset? What is the resulting training error? (Because there might be different reasonable answers, make sure to justify your response. Note that the training error in this question is the fraction of mis-classified vectors in the training data.)

**(b)** Are there some values of $K$ that might be too large or too small for this dataset?

**(c)** Please sketch the 1-nearest neighbor decision boundary for this dataset.

**Question 3** (LASSO)
Consider a vector $\beta$ observed through noisy measurements $y$,

$$y = X\beta + z. \tag{1}$$

Our goal is to recover or estimate $\beta \in \mathbb{R}^N$, given $X \in \mathbb{R}^{M\times N}$ and $y \in \mathbb{R}^M$. Below, you will show in several steps that when $\beta$ and $z$ are modeled as independent and identically distributed (i.i.d.) Gaussian, maximum a posteriori (MAP) estimation of $\beta$,

$$\beta_{MAP} = \arg\max_{\beta} f\left(\beta|X, y\right),$$

is a special case of the least absolute shrinkage and selection operator (LASSO). (Note that this problem is closely related to a problem from the 2019 final exam, and some parts that students had to explain in that exam are not required on this one.)
We begin deriving the solution $\beta_{MAP}$ that maximizes $f\left(\beta|X, y\right)$,

$$\beta_{MAP} = \arg\max_{\beta} f\left(\beta|X, y\right) = \arg\max_{\beta} \frac{f(\beta, X, y)}{f(X, y)} = \arg\max_{\beta} f(\beta, X, y).$$

Focusing on the last term, $f(\beta, X, y) = f(X)f(\beta|X)f(y|\beta, X)$, but $\beta$ is independent of $X$, i.e., $f(\beta|X) = f(\beta)$, and so

$$\beta_{MAP} = \arg\max_{\beta} f(\beta, X, y) = \arg\max_{\beta} f(X)f(\beta)f(y|\beta, X) = \arg\max_{\beta} f(\beta)f(y|\beta, X). \tag{2}$$

**(a)** To compute $f(\beta)$ and $f(y|\beta, X)$ in (2), we model each of the $M$ scalar entries in $z \in \mathbb{R}^M$ (1), which can be interpreted as a noise or error vector, as i.i.d. zero-mean Gaussian with variance $\sigma_Z^2$. That is, $Z_m \sim \mathcal{N}(0, \sigma_Z^2)$, where $Z_m$ is the random variable corresponding to entry $m$ of the noise vector, and $m \in \{1, \ldots, M\}$. The pdf for $Z_m$ can be expressed,

$$f(Z_m = z) = \frac{1}{\sqrt{2\pi\sigma_Z^2}} e^{-\frac{z^2}{2\sigma_Z^2}}.$$

You are also given that the $N$ entries of $\beta$ are i.i.d. with pdf

$$f(\beta_n) = c_1 e^{c_2|\beta_n|},$$

where $n \in \{1, \ldots, N\}$, and $c_1 > 0$ and $c_2 < 0$ are constants that control the variance of this RV. Derive expressions for $f(\beta)$ and $f(y|\beta, X)$ in terms of $X, y, \sigma_z, c_1, c_2$. (Hint: you can simplify your expressions using $\ell_1$ and $\ell_2$ norm notations for $\beta$ and $y - X\beta$.)

**(b)** Because the logarithm is a monotone function, it suffices to maximize $\log\left(f\left(\beta|y, X\right)\right)$,

$$\begin{aligned}
\beta_{MAP} &= \arg\max_{\beta} f(\beta)f(y|\beta, X) \\
&= \arg\max_{\beta} \log(f(\beta)f(y|\beta, X)) \\
&= \arg\max_{\beta} \text{Polynomial}(\beta, y, X, \sigma_Z, c_1, c_2).
\end{aligned}$$

Express Polynomial$(\beta, y, X, \sigma_Z, \sigma)$.

**(c)** The LASSO has the form

$$\beta_{\text{LASSO}} = \arg\min_{\beta} \|y - X\beta\|^2 + \lambda\|\beta\|_1,$$

where $\|\cdot\|^2$ is the squared $\ell_2$ norm, and $\|\cdot\|_1$ is the $\ell_1$ norm. The LASSO form should correspond to your expression above; what is $\lambda$? What intuition can you draw from the expression for $\lambda$? If you could not derive the expression for $\beta_{MAP}$, please explain how $\lambda$ impacts the LASSO solution.