ECE523: Engineering Applications of Machine Learning and Data Analytics

I acknowledge that this exam is solely my effort. I have done this work by myself. I have not consulted with others about this exam in any way. I have not received outside aid (outside of my own brain) on this exam. I understand that violation of these rules contradicts the class policy on academic integrity.

Name: Solution

Signature: ____________________

Date: ____________________

Instructions: There are five problems. You have 50 minutes to complete the exam. Partial credit is given for answers that are partially correct. No credit is given for answers that are wrong or illegible. Write neatly.

Problem 1: __________

Problem 2: __________

Problem 3: __________

Problem 4: __________

Problem 5: __________

Total: __________

Problem #1 – Ridge Regression (10 Points)

In class we discussed linear discriminant models and one approach was linear regression. In this problem we look at ridge regression, which is given by

$$\arg\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$$

where $\mathbf{y} \in \mathbb{R}^n$ is a vector of the outputs, $\mathbf{X} \in \mathbb{R}^{n\times p}$ is the matrix of data and $\mathbf{w} \in \mathbb{R}^p$ are the parameters for the linear model $y = \mathbf{w}^\mathsf{T}\mathbf{x}$. Find $\mathbf{w}$.

Solution

This problem is very similar to the linear regression problem that we discussed in class; however, now we have a new regularization term on the set of parameters. The parameter λ is defined by the user so it will end up in the solution for $\mathbf{w}$. The good news is that this problem can be solved the same was a linear regression. The only difference is that we end up with a slightly different solution. We begin by re-writing the optimization problem as:

$$L(\mathbf{w}) = \frac{1}{2}(\mathbf{y}-\mathbf{X}\mathbf{w})^\mathsf{T}(\mathbf{y}-\mathbf{X}\mathbf{w}) + \frac{\lambda}{2}\mathbf{w}^\mathsf{T}\mathbf{w}$$

Now we use standard calculus to solve the problem.

$$\begin{aligned}
\frac{\mathsf{d}L}{\mathsf{d}\mathbf{w}} &= -\mathbf{X}^\mathsf{T}(\mathbf{y}-\mathbf{X}\mathbf{w}) + \lambda\mathbf{w} = 0 \\
-\mathbf{X}^\mathsf{T}\mathbf{y} + \mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w} + \lambda\mathbf{w} &= 0 \\
\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w} + \lambda\mathbf{w} &= \mathbf{X}^\mathsf{T}\mathbf{y} \\
(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda I)\mathbf{w} &= \mathbf{X}^\mathsf{T}\mathbf{y} \\
\mathbf{w} &= (\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda I)^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}
\end{aligned}$$

Problem #2 – Principal Component Analysis (10 Points)

In class, we showed two different approaches that we could arrive at a solution to PCA: one with linear algebra and one with optimization. This problem asks you to use both that you know about the PCA projection and task of optimization. Use these facts:

- The projection is performed with $z = \mathbf{w}^{\mathsf{T}}\mathbf{x}$. Note that z is a scalar because we are only looking for one principal axis.
- I am not too concerned with the magnitude of $\mathbf{w}$, but I am concerned with its direction.
- You need to maximize the variance of z.

Use these facts to find $\mathbf{w}$. It maybe a good idea to let $\mathbf{X} \in \mathbb{R}^{p \times n}$ be the matrix of data. Then the covariance matrix is given by $\frac{1}{n-1}\mathbf{X}\mathbf{X}^{\mathsf{T}} = \Sigma$. This approach is similar to how we discussed PCA from a linear algebra perspective.

Solution

This one is verbatim from your text book (and the in class notes would have been sufficient too)! The projection is performed with $z = \mathbf{w}^{\mathsf{T}}\mathbf{x}$ and we know that $\text{Var}(X) = \frac{1}{n-1}\mathbf{X}\mathbf{X}^{\mathsf{T}} = \Sigma$. Then we have

$$\text{Var}(z) = \mathbf{w}^{\mathsf{T}}\Sigma\mathbf{w}$$

We seek to fin the $\mathbf{w}$ such that $\text{Var}(z)$ is maximized as asked in the question; however, we have the constraint that $\|\mathbf{w}\|_2^2 = 1$[1]. As we saw with other problems we've ecountered in the homework and lecture, this is a constrained optimization problem where we want to maximize $\mathbf{w}^{\mathsf{T}}\Sigma\mathbf{w}$ subject to $\|\mathbf{w}\|_2^2 = 1$ Writing this in the form of the Lagrangian gives us

$$L(\mathbf{w}, \eta) = \mathbf{w}^{\mathsf{T}}\Sigma\mathbf{w} - \eta(\|\mathbf{w}\|_2^2 - 1)$$

which is exactly what we had in class! Taking the derivative w.r.t. $\mathbf{w}$, we find that $\Sigma\mathbf{w} = \eta\mathbf{w}$, which means that $\mathbf{w}$ is an eigenvector of Σ.

[1] You could even assume that $\|\mathbf{w}\|_2^2 = d$, where $d > 0$.

Problem #3 – A Gamblers Ruin (10 Points)

[True/False] (1 point): Density estimation (using say, the kernel density estimator) can be used to perform classification.
Solution: The correct answer to this question can be addressed via kernel density estimation or k-NN classifiers. In regards to kernel density estimation, we can estimate the quantity $p(\mathbf{x}|\omega)$ then use it directly with

$$\omega^* = \arg\max_{\omega \in \Omega} p(\mathbf{x}|\omega)P(\omega)$$

Thus, the answer to this question is True.

[True/False] (1 point): One of the disadvantages of the logistic function is that its derivative is not very convenient to compute.
Solution: Given a logistic function $f(x)$, we know from the lecture that the derivative of f is $f'(x) = f(x)(1 - f(x))$. Thus, the derivative can be written in terms of the function evaluation itself, which is very easy to compute. Hence, the answer to this question is False.

[True/False] (1 point): Logistic regression assumes that the log-likelihood ratio for two classes with equal priors is linear. More formally this is given by

$$\log \left\{ \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} \right\} = \mathbf{w}^\mathsf{T}\mathbf{x} + w_0$$

Solution: This question was taken directly from the lecture notes and the book. The answer to this question is True.

[True/False] (1 point): Regularization is one way to prevent overfitting and the reason it is so effective is because the regularization term is data-dependent. Therefore, the optimization process will "find" the best way to be resilient against overfitting.
Solution: There are two things going on in this statement. The first is that regularization is on way to prevent overfitting, which is True; however, regularization is not a function of the data rather one of the parameters. Thus, regularization is data independent, which makes the overall statement False.

[True/False] (1 point): The training error of 1-NN classifier is 0.
Solution: Each point is its own neighbor, so 1-NN classifier achieves perfect classification on training data.

[True/False] (1 point): The principal components are the ones that maximize the variance within a class.
Solution: This is not a true statement since we know that PCA does not account for variations within a class; rather the entire data set.

[True/False] (1 point): The correspondence between logistic regression and Gaussian naïve Bayes (with identity class covariances) means that there is a one-to-one correspondence between the parameters of the two classifiers.
Solution: Each logistic regression model parameter corresponds to a whole set of possible Gaussian naïve Bayes classifier parameters, there is no one-to-one correspondence because logistic regression is discriminative and therefore doesn't model $P(X)$, while GNB does model $P(X)$.

[True/False] (1 point): The number of actions need not be equal to the number of classes.
Solution: Clearly this statement is not true because we could always take an action such as fail to classify a data point.

[True/False] (1 point): I don't like true and false questions, but I do like free points!
Solution: To each their own.

[Accept/Reject] (1 point): "My algorithm is better than yours. Look at the training error rates!"
Solution: I would lean to reject this manuscript because they are making the statement "better" based on the training error.

[Accept/Reject] (1 point): "My algorithm is better than yours. Look at the training error rates and the p-value from the signed rank Wilcoxon test! (Footnote: reported results for best value of λ, chosen with 10-fold cross validation.)"
Solution: I would still lean to reject this manuscript because they are making the statement "better" based on the training error.

Problem #4 – To Bayes or Not Bayes (10 Points)

Let consider a Bayes classifier with $p(\mathbf{x}|\omega_i)$ distributed as a multivariate Gaussian with mean μ_i and covariance $\Sigma_i = \sigma^2 I$ (note they all share the same covariance). We choose the class that has the largest

$$g_i(\mathbf{x}) = \log(p(\mathbf{x}|\omega_i)P(\omega_i)) \propto \mathbf{w}_i^\mathsf{T}\mathbf{x} + w_{0i}$$

Find $\mathbf{w}_i$ and w_{0i}. Fact:

$$p(\mathbf{x}|\omega_i) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_i|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\left(\mathbf{x}-\mu_i\right)^\mathsf{T}\Sigma_i^{-1}\left(\mathbf{x}-\mu_i\right)\right\}$$

Hints: start $g_i(\mathbf{x})$ and the fact stated above. Then begin to drop out the terms that are constant for all $g_i(\mathbf{x})$.

Solution

We begin by using the definition of the likelihood of a multi-variate a Gaussian then beginning to reduce down the expression by removing terms that do not change over $i \in [c]$ where c is the number of classes. We are told to choose class with the largest $g_i(\mathbf{x})$, or

$$\begin{aligned}
\arg\max_{i\in[c]} g_i(\mathbf{x}) &= \arg\max_{i\in[c]}\left\{-\frac{1}{2}\left(\mathbf{x}-\mu_i\right)^\mathsf{T}\Sigma_i^{-1}\left(\mathbf{x}-\mu_i\right) - \frac{d}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma_i|) + \log(P(\omega_i))\right\} \\
&= \arg\max_{i\in[c]}\left\{-\frac{1}{2\sigma^2}\left(\mathbf{x}-\mu_i\right)^\mathsf{T}\left(\mathbf{x}-\mu_i\right) + \log(P(\omega_i))\right\} \\
&= \arg\max_{i\in[c]}\left\{-\frac{1}{2\sigma^2}\|\mathbf{x}-\mu_i\|_2^2 + \log(P(\omega_i))\right\}
\end{aligned}$$

where the second step from the terms that are constant for all $i \in [c]$. We can further reduce the expression by expanding out $\|\mathbf{x}-\mu_i\|_2^2$, which gives us

$$\begin{aligned}
\arg\max_{i\in[c]}\{g_i(\mathbf{x})\} &= \arg\max_{i\in[c]}\left\{-\frac{1}{2\sigma^2}\left(\mathbf{x}-\mu_i\right)^\mathsf{T}\left(\mathbf{x}-\mu_i\right) + \log(P(\omega_i))\right\} \\
&= \arg\max_{i\in[c]}\left\{-\frac{1}{2\sigma^2}\left(\mathbf{x}^\mathsf{T}\mathbf{x} - 2\mu_i^\mathsf{T}\mathbf{x} + \mu_i^\mathsf{T}\mu_i\right) + \log(P(\omega_i))\right\} \\
&\arg\max_{i\in[c]}\left\{-\frac{1}{2\sigma^2}\left(-2\mu_i^\mathsf{T}\mathbf{x} + \mu_i^\mathsf{T}\mu_i\right) + \log(P(\omega_i))\right\} \\
&= \arg\max_{i\in[c]}\left\{\underbrace{\frac{1}{\sigma^2}\mu_i^\mathsf{T}}_{\mathbf{w}_i}\mathbf{x} + \underbrace{\left(\log(P(\omega_i)) - \frac{1}{2\sigma^2}\mu_i^\mathsf{T}\mu_i\right)}_{w_{0i}}\right\}
\end{aligned}$$

Problem #5 – Density Estimation (10 Points)

In class, we discussed three conditions that need be met if a density estimator ($p_n(\mathbf{x}) = \frac{k_n/n}{V_n}$) is to converge in probability to the true density ($p(\mathbf{x})$). More formally,

$$\lim_{n\to\infty} V_n = 0, \quad \lim_{n\to\infty} k_n = \infty, \quad \lim_{n\to\infty} \frac{k_n}{n} = 0$$

where k_n is the number of samples that fall within a region $\mathcal{R}$ with volume V_n. Describe two out of the three conditions and why they are necessary for $p_n(\mathbf{x})$ to converge in probability to $p(\mathbf{x})$ when n approaches infinity.

Solution

If $p_n(\mathbf{x})$ to converge in probability to $p(\mathbf{x})$ then we must have these three conditions be met. The assures us that the space averaged by P/V_n (see lecture notes) will converge to $p(\mathbf{x})$ provided that the regions shrink uniformly. The second condition is there if $p(\mathbf{x}) = 0$, which assures use that the frequency ratio will also converge. Finally, the last condition need to be held if there is any type of convergence.