CMPUT 466/566
University of Alberta **Fall 2022**

Assignment 2

Due November 17, 2022, 11:59 pm on eClass
566 students: 77 marks total
466 students: 67 marks total

This assignment contains a written portion, and a programming portion. Please hand in a pdf for your written answers, and a zip of your code for the programming portion. Your pdf can contain scans of handwritten answers, so long as they are legible.

Questions

1) [Interpretability/Ethics] (12 marks)

For this sections you need to use explainer.ipynb to explain machine learning model predictions.

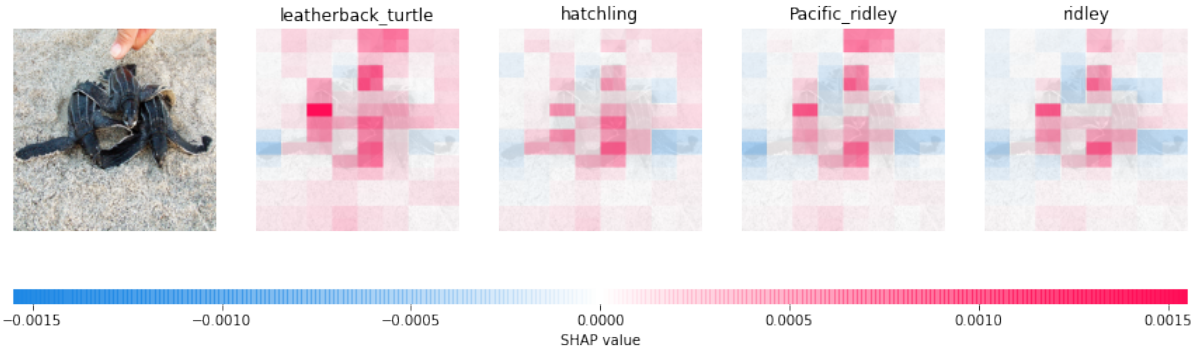

Figure 1: Interpretability example

Figure 1 is an example of image explainability plot. The original image is featured on the left. The following 4 images show some explainability information about the 4 classes with the highest probability determined by the model. These annotated images are obtained by manipulating part of the image and rerunning classification. The red parts of the images indicate that if that part of the image is modified then the probability of the class decreases (that part of the image increases the probability that class is predicted). The blue parts of the image indicate that if that part of the image is modified then the probability of the class increases (that part of the image decreases the probability that class is predicted). You can use this to try to determine why the classifier is making its predictions.

Run the explainer.ipynb to answer the following question

(a) (2 marks) What is the most important part of the image for determining a *President of the United States* in the image of Bush and Obama?

(b) (2 marks) Why is the face mask misclassified as *Domino/Eye mask*?

(c) (4 marks) Find another example of an image that is misclassified by the classifier? You can use the *load_image* function to load and crop images. Why is it misclassified? (Include a copy of the explainability plot in your PDF)

(d) I(2 marks) mage classification models, like most other machine learning models have unwanted bias. Identify one source of potentially unwanted bias in the model by interpreting the prediction on the image of French President Emmanuel Macron and German Chancellor Angela Merkel.

(e) (2 marks) Give an example of an ethical implication that can come from the choice of classes to use during training.

2) [Support Vector Machines:] (25 marks)

Let us focus on a non-linearly separable dataset $\{(x_i, y_i)|i = 1, \cdots, m\}$ where $y_i \in \{+1, -1\}$ is the class label. We can use a support vector machine (SVM) with soft margin formulation to solve this classification task. We introduce slack variable ξ_i, the margin error coefficient C, and Lagrangian multipliers α_i. In this question, we use kernels $k(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ which project data points into high dimensional space. Also, we assume the interception $b = 0$ for simplicity. Then, the prediction is given by the sign of function $f(x_i) = w^T \phi(x_i)$.

(a) (10 marks) Basic facts of SVMs. Please answer True or False.

i. Data points with Lagrangian multipliers set to zero do not contribute to the prediction. Why or why not? [Hint: Write out the learnt prediction function.]

ii. If we require slack variables to be larger than zero, no support vectors can be misclassified. Why or why not?

iii. If a dataset is linearly separable, after training a hard margin SVM, there will be no points between the hyperplanes defined by y=1 and y=-1. Why or why not?

iv. The distance between the separation plane and the closest point equals to

$$\min_i \frac{y_i w^T \phi(x_i)}{\|w\|}.$$

Why or why not?

v. Usually SVM maximizes the margin between $f(x) = +1$ and $f(x) = -1$. Let's say instead that we want to maximize the margin between $f(x) = +2$ and $f(x) = -2$. The newly learnt model parameters do not change the distance between the separation plane and the closest point. Why or why not?

(b) (5 marks) Let us focus on a soft-margin SVM with a Gaussian kernel $k(x, z) = e^{-\gamma \|x-z\|^2}$.

i. Figure 2 shows different learnt separation planes with different kernels: linear kernel, second-order polynomial kernel and fifth-order polynomial kernel. Could you match the results in the picture with the kernel used? Does high kernel dimension or low kernel dimension more possibly cause overfitting?

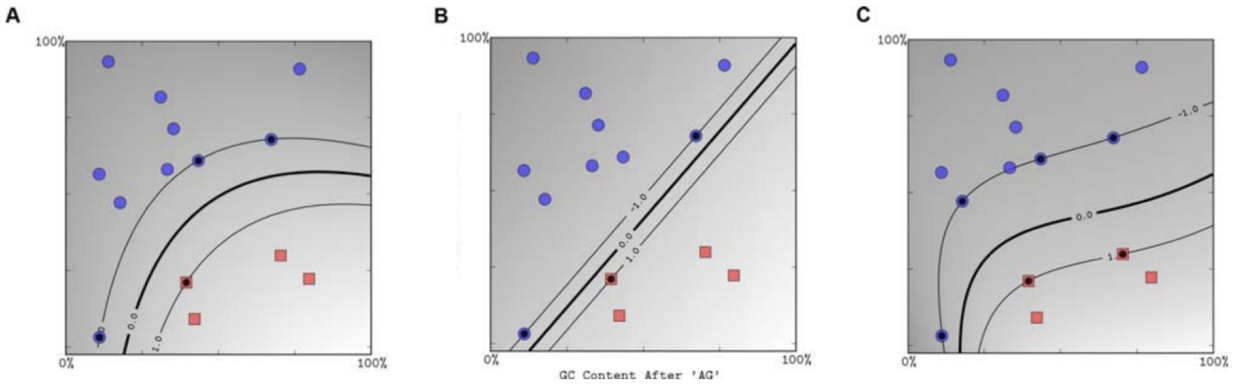

Figure 2: Separation Planes with Different Kernels. For the purposes of this question you can ignore the axis labels.

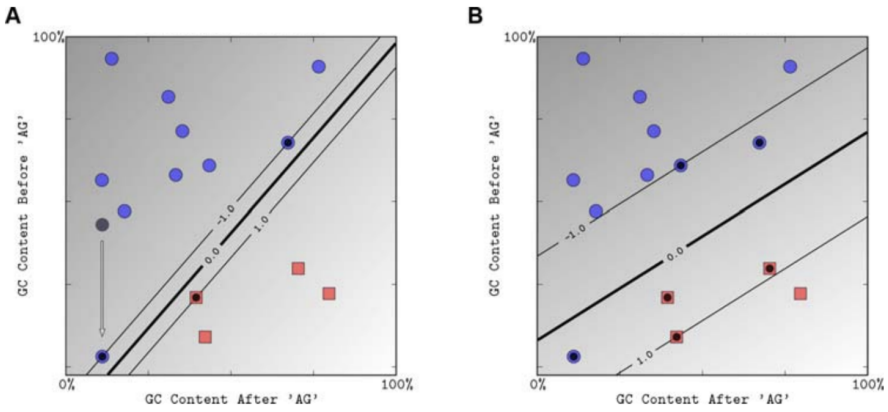

Figure 3: Separation Planes with Different *C* Values. For the purposes of this question you can ignore the axis labels.

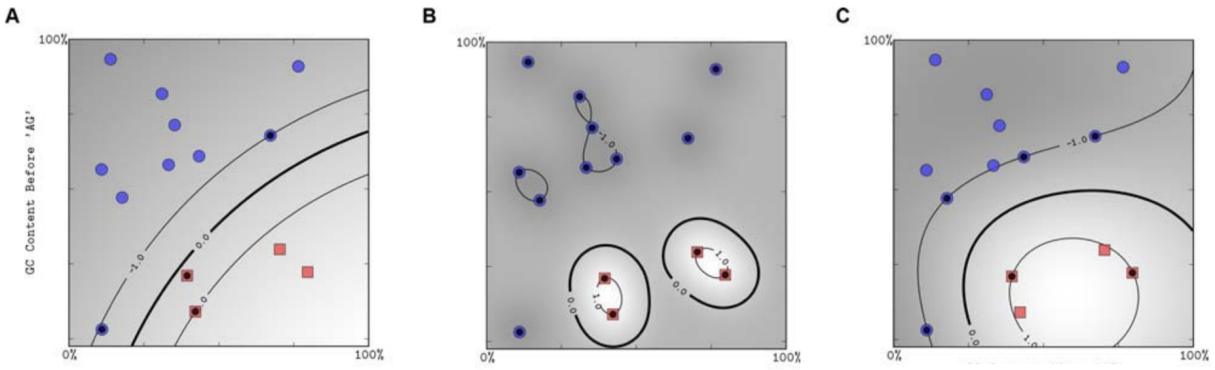

Figure 4: Separation Planes with Different γ Values. For the purposes of this question you can ignore the axis labels.

ii. Figure 3 shows different learnt separation planes with different C values: 2 and 200. Match the results in the picture with the value used. Does large C or small C cause overfitting in this case?

iii. Figure 4 shows different separation planes learned with different kernel widths γ: 0.0025, 1, and 400 but a fixed C value. Match the results in the picture with the value used. With a fixed sufficiently large C, does large γ or small γ cause overfitting in this case?

(c) (10 marks) Let us focus on a soft-margin SVM with Gaussian kernels $k(x,z) = e^{-\gamma\|x-z\|^2}$. Consider a dataset where the within-class and between-class distances for any two points are bounded. That is, for any two points x_i and x_j in the same class, their Euclidean distance is upper bounded, that is $\|x_i - x_j\| \leq s_1$ and for any two points x_i and x_j in different classes, their distance is lower bounded, that is $\|x_i - x_j\| \geq s_2$. Also, $s_1 > s_2$. In this question, we would like to find values of the hyperparameters C and γ such that there will not be any misclassified data outside the margin and the margin is the area between the planes $f(x) = +1$ and $f(x) = -1$.

i. Let us separate the data into two sets where $\mathscr{S}_i = \{j | y_j = y_i\}$ contains data in the same class as x_i and $\mathscr{D}_i = \{j | y_j = -y_i\}$ contains data in the different class from x_i. Prove that for all i,

$$y_i(f(x_i) - y_i) \geq \sum_{j \in \mathscr{S}_i} \alpha_j e^{-\gamma s_1^2} - \sum_{j \in \mathscr{D}_i} \alpha_j e^{-\gamma s_2^2} - 1.$$

ii. Continue working on the above inequality and prove that for all i,

$$y_i(f(x_i) - y_i) \geq Cm\gamma(s_2^2 - s_1^2) - 1.$$

[Hint: think about the properties of Lagrangian multipliers. Also, you may need the following property for convex functions: for any convex function f, we have

$$f(x) \geq f(y) + f'(y)(x - y).]$$

iii. Finally, find a bound for the product of hyperparameters $C\gamma$ so there will be no misclassified data outside the margin. [Hint: consider what condition is needed for no misclassified data outside the margin.]

3) [Convolutional neural networks] (25 marks)

For this question, please make a copy of this notebook and save it on your Google Drive. The datasets used for this question are linked here. We highly recommend using Google Colaboratory for this question to avoid longer training times and compatibility issues on your local machine. Moreover, to access the datasets in Colab, it is recommended that you link your Google Drive to your notebook.
For submitting the solutions, please download the .ipynb version of your file with the cell outputs by clicking *File > Download*. You should also submit a .pdf file of our notebook after running all the cells. Make sure the pdf has outputs of all the cells including your answers to theoretical questions. To download the notebook as a pdf go to *File > Print*. Please keep the same name for your notebook and pdf file, for example CNN.ipynb and CNN.pdf.

4) [Adversarial Examples in Machine Learning] (466 students: 5 marks; 566 students: 15 marks)

White Box Attack (FGSM) Let us consider the white box adversarial attack FGSM (Fast Gradient Sign Method). Given a loss function J, input x and neural network weights w , the attack is formulated as follows:

$$x = x_0 + \eta \cdot \text{sign}(\nabla_x J(x_0; w)). \tag{1}$$

(a) (**566 students only**, 10 marks) The basic concept of this kind of adversarial attack is that we perturb the input x by r such that the loss function J is maximized. However in such a case, r needs to be bounded. Because we want the perturbations to be almost imperceptible to humans, we can apply some condition like $\|r\|_p \leq \eta$.

i. Frame the above concept as an optimization problem with the constraint being $\|r\|_\infty \leq \eta$
Note: $x = x_0 + r$

ii. Solve the optimization problem and prove that this leads to the attack being Equation (1).
Hints and Steps you may follow:

- Use first order Taylor approximation on the cost function J to get something like $\nabla_x J(x_0; w)$
- Get an upper bound of the previous expression using Holder's inequality:

$$\begin{aligned} &\langle x\,,\, y\rangle \leq \|x\|_p \|y\|_q \\ \text{where } &\frac{1}{p} + \frac{1}{q} = 1 \end{aligned} \tag{2}$$

Note that for a function f, $\max(f) \leq \max(\text{Upper Bound of } f)$

- Use the given constraint to get another upper bound

- Note that the L1 norm is the sum of absolute values
- This fact may be helpful at this stage: $|a| = a \cdot sign(a)$
- The simplified form should be independent of r. We may conclude here that this form is the maximum value of the cost and find the value of r from here. Note: You do not have to consider the tightness of the bounds.

(b) (**466 and 566 students**, 5 marks) Consider the following neural network to predict y :

$$\begin{aligned} z &= W_1 x + b \\ y &= W_2 z \\ x &\in \mathbb{R}^p,\ y \in \mathbb{R}^q \end{aligned} \tag{3}$$

Consider that the loss function J is the L2 loss between the output y and target $\hat{y}$. For a given input x_0, compute the perturbed input under FGSM attack.