# midterm

October 15, 2024

## 1 STP598 Machine Learning & Deep Learning

### 1.1 Midterm Exam (Take-home)

#### 1.1.1 Due 11:59pm Friday Oct. 25, 2024 on Canvas

#### 1.1.2 name, id

### 1.2 Question 1

- In multiple linear regression, we have residual vector defined as $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\beta}$. Prove that it is perpendicular to the column space of $\mathbf{X}$, i.e. $\mathbf{X}^T\mathbf{e} = \mathbf{0}$.
- Now if we want to plot $\hat{\mathbf{y}}$ against $\mathbf{y}$, what is the slope? Can you prove it?

### 1.3 Question 2

*Probit regression* is a binary classification model alternative to logistic regression. The link function is probit function (inverse CDF of standard normal $\Phi^{-1}$) instead of logit function, i.e.

$$\Phi^{-1}(\Pr(Y = 1|X)) = X\beta$$

- Write down the log-likelihood function of $\{x_i, y_i\}_{i=1}^n$ and answer briefly how you can find the solution $\hat{\beta}$.
- Fit **digits** data (using `sklearn.datasets.load_digits` for `n_class=2`) with *logistic regression*, *probit regression*, *random forest*, and *Gaussian process classifier* respectively. Split and dataset into training and testing (e.g. 80% vs 20%). Compare their testing accuracy in one table. Hint: probit regression is not implemented in `scikit-learn` but has been implememted in `statsmodels`. Consider statsmodels.discrete.discrete_model.Probit

### 1.4 Question 3

Compare impurity measures for splitting nodes in trees.

- Fill in the blanks of the table to compute Gini index, Shannon entropy and misclassification error.

| | Class 1 | Class 2 | Class 3 | $\hat{p}_1$ | $\hat{p}_2$ | $\hat{p}_3$ | Gini | Entropy | Error |
|---|---|---|---|---|---|---|---|---|---|
| $\mathcal{A}$ | 3 | 3 | 4 | | | | | | |
| $\mathcal{A}_L$ | 1 | 0 | 3 | | | | | | |

| | Class 1 | Class 2 | Class 3 | $\hat{p}_1$ | $\hat{p}_2$ | $\hat{p}_3$ | Gini | Entropy | Error |
|---|---|---|---|---|---|---|---|---|---|
| $\mathcal{A}_R$ | 2 | | 3 | | 1 | | | | |

- Compute the impurity reductions for the **three** measures.

### 1.5 Question 4

Gaussian process is a flexible tool for modeling nonlinear functional relationship. Given data $\{x_i, y_i\}_{i=1}^n$, we assume the following model:

$$y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$$
$$f \sim \mathcal{GP}(0, \mathcal{C})$$

- Given a new location $x_*$, predict $\hat{y} = f(x_*)$ and give the uncertainty estimate (credible interval).
- Simulate a dataset of 1-d input $x$ and output $y$, e.g. using $y = \sin(x) + .1 * N(0, 1)$, for $n_1 = 10$ points. Use Gaussian process to fit such dataset. Predict $x_*$ on a grid of 100 points over the defined domain ($[0, \pi]$ for example). Now increase the data to $n_2 = 50$ points (may contain $n_1$ points), repeat the same prediction. Plot the following on the same graph:
  - $n_1$ data points and $n_2$ data points with different colors (scatter plot)
  - posterior prediction lines based on $n_1$ and $n_2$ respectively with different colors (line plot).
  - posterior credible bands based on $n_1$ and $n_2$ respectively with different colors (fill_between)
- Compare the plots between two cases ($n_1$ vs $n_2$). What do you find?

### 1.6 Extra*

Please comment on this course. What suggestions do you have to improve this course?