DS-GA 1003 Machine Learning: Homework 1

Due 11.59 p.m. EST, February 27, 2024 on Gradescope

(fill in your name here)
(collaborators if any)

We encourage LaTeX-typeset submissions but will accept quality scans of hand-written pages.

1 Linear Regression Model

Consider the data generating process as such: $\mathbf{x} \in \mathbb{R}^D$ is drawn from some unknown $p(\mathbf{x})$ and $y = w_1^{true} x_1 + \epsilon_y$, where $w_1^{true} \in \mathbb{R}$ and $\epsilon_y \sim \mathcal{N}(0, 1)$. This is unknown to us, as a result, we construct a linear model for y using all D features of $\mathbf{x}$, instead of just using x_1.

(A) Explain what the terms **model class** and **model misspecification** mean. Is our model correctly *specified* here? Why or why not?

Solution. Write your solution for each question using the `solution` environment. Feel free to use style packages to your convenience, e.g. highlighting parts of your solution that you still need to work on. □

Let $\widehat{w_1}$ be the estimate of w_1^{true} using only x_1 and let $\widehat{w_1^{all}}$ be the estimate of w_1^{true} when using all of $\mathbf{x}$. We will study the effects of our model by analyzing the relationships between $\mathbb{E}[\widehat{w_1^{all}}]$ and $\mathbb{E}[\widehat{w_1}]$, as well as between $\text{Var}[\widehat{w_1^{all}}]$ and $\text{Var}[\widehat{w_1}]$. We do so empirically by running PyTorch simulations as follows:

1. Pick any value of w_1^{true} you like as ground truth, e.g. with `torch.randn(1)`.
2. Write a function, taking D and c as input, that does the following: **(1)** Generate $N = 50$ samples of $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \Sigma)$, where Σ is the $D \times D$ covariance matrix with all diagonal entries equal to $\sigma^2 = 1$ and all off-diagonal entries equal to c. **(2)** Compute y using the relationship above (note that y only depends on the first feature). This involves drawing N samples of noise $\epsilon_y \sim \mathcal{N}(0, 1)$. **(3)** Using our dataset of N samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, compute the least-squares solutions for $\widehat{w_1^{all}}$ (i.e. using all features and taking the first coefficient) and $\widehat{w_1}$ (i.e. using only one feature).
3. Write a function that performs Step 2 for $T = 100$ trials, i.e. each trial generates a new dataset to compute $\widehat{w_1^{all}}$ and $\widehat{w_1}$. (Note that w_1^{true} is constant throughout.) For each estimator, compute the mean and standard deviation of the T trials.
4. Perform Step 3 for each $c \in \{0.1, 0.2, 0.3, \ldots, 0.9\}$ and each $D \in \{2, 4, 8, 16, 32\}$. Separately plot the means and standard deviations as a function of c, using the same plot for both estimators. This means you should have 10 plots altogether: two plots (means and standard deviations) for each of the five choices of D. Each plot will contain two curves (the two estimators).

(B) What do you observe with respect to c and D? How do you explain your results? Show a few (not all) of your 10 generated plots to support your answer.

2 Bayesian Linear Regression Model

Consider the data generating process as such: $x \sim \mathcal{N}(0, 1)$ and $y = w_{true}x^2 + \epsilon$, where $w_{true} = 1.0$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$, and $\sigma^2 = 1.0$. Again, this is unknown to us. We will model the data using Bayesian linear regression.

(A) Using PyTorch, simulate a dataset $\mathcal{D}_N = \{(x_i, y_i)\}_{i=1}^N$ for each $N \in \{10, 100, 1000, 10000\}$ according to the true data generating process above. For our Bayesian linear regression model, let us choose our prior as $w \sim N(0, 1)$ and our likelihood as $y|w, x \sim N(wx, \sigma^2)$ for $\sigma^2 = 1.0$. Compute the mean and variance of the posterior $w|\mathcal{D}_N$ for each dataset. Does the posterior concentrate on w_{true}? Why or why not?

(B) What would be challenging about our analysis in part (A) if we had picked a different prior, for example, Laplace or Gamma?

(C) Repeat part (A), except we use the basis set $\boldsymbol{\phi}(x) = [x, x^2]$ (instead of x itself) and perform 2D Bayesian linear regression. We choose our prior to be $\mathbf{w} \sim N(\mathbf{0}, \mathbf{I})$, where $\mathbf{I}$ is the 2×2 identity matrix, and our likelihood to be $y|\mathbf{w}, x \sim N(\mathbf{w}^\top \boldsymbol{\phi}(x), \sigma^2)$ for $\sigma^2 = 1.0$. Compute the mean and variance of the posterior $\mathbf{w}|\mathcal{D}_N$ for each dataset. What do you observe about the posterior as N changes? Why?

(D) Reflecting on your answers in parts (A) and (C), name one challenge that **cannot be solved** by using a Bayesian model (instead of a frequentist approach like standard linear regression).

(E) Reflecting on your answers in part (C) and in Question 1, name one challenge that **can be improved** by using a Bayesian model.

Hint: Both part (C) above and Question 1 involve doing linear regression with many correlated features. How do these two sets of findings relate? Does using a Bayesian approach affect the way the model treats correlated features?

3 Class-Conditional Gaussian Generative Model

Consider a classification task where $\mathbf{x} \in \mathbb{R}^D$ and $y \in \{1, \ldots, K\}$. We observe the dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$. Let us construct a model for the joint distribution as

$$p_\theta(\mathbf{x}, y) = p_\theta(\mathbf{x}|y)p_\theta(y)$$

where θ denotes the set of all parameters of the model.

(A) Our model is known as a **class-conditional generative model**. What about the model makes it generative? What makes it class-conditional?

(B) For a given value of θ, how would you predict the label for a new test point $\mathbf{x}_\star$ using your model $p_\theta(\mathbf{x}, y)$?

Let us model y as a Categorical distribution $\text{Cat}(\boldsymbol{\pi})$. Here $p_\theta(y = k) \triangleq \pi_k$, where $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_K]$ such that $\forall k$, $\pi_k \geq 0$ and $\sum_k \pi_k = 1$. You may leave $p_\theta(\mathbf{x}|y)$ unspecified for now.

(C) Write down an expression for the log-likelihood of the observed dataset $\mathcal{D}$.

(D) Derive an expression for the maximum likelihood estimator (MLE) for $\boldsymbol{\pi}$, which we will denote as $\widehat{\boldsymbol{\pi}}$. Make sure to account for the constraints on $\boldsymbol{\pi}$ using Lagrange multipliers.

Let us further model $\mathbf{x}|y$ as (multivariate) Gaussian distributions $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \Sigma_k)$ for all K classes, where $\boldsymbol{\mu}_k \in \mathbb{R}^D$ and Σ_k is a $D \times D$ (positive semi-definite) covariance matrix. Assume that there are only $K = 2$ classes. This means that the total set of parameters are $\theta = \{\boldsymbol{\pi}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2\}$.

Now, consider the case where the data comes from this model. That is, $y \sim \text{Cat}(\boldsymbol{\pi}^{true})$ and $\mathbf{x}|y = k \sim \mathcal{N}(\boldsymbol{\mu}_k^{true}, \Sigma_k^{true})$ for all k. After we observe this data, we can then construct a *discriminative model* to predict y from $\mathbf{x}$, that is, we will learn a model for $p_{true}(y = k|\mathbf{x})$. Let us do so using **logistic regression**:

$$y|\mathbf{x} \sim \text{Bernoulli}(\sigma(\mathbf{w}^\top \mathbf{x}))$$

where $\mathbf{w}$ are the parameters of the model and $\sigma(z)$ is the logistic sigmoid:

$$\sigma(z) = \frac{1}{1 + \exp[-z]}$$

(E) Will logistic regression always be able to model the true data conditional $p_{true}(y = k|\mathbf{x})$? If so, why? If sometimes, when? And if there are any cases where logistic regression will not be able to model $p_{true}(y = k|\mathbf{x})$, are there any ways to fix it?

4 Poisson Generalized Linear Model

Consider a classification task where $\mathbf{x} \in \mathbb{R}^D$ and $y \in \mathbb{N} = \{0, 1, 2, 3, \dots\}$, noting that the support of y is the unbounded set of natural numbers. We have an observed dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$. Let us also assume that the number of features, D, is larger than the number of examples, N. We will model this data using a Poisson Generalized Linear Model (GLM). Let $\boldsymbol{\theta}$ denote the linear coefficients of the model.

(A) Write down the log-likelihood function of the Poisson GLM.

(B) Given a test point $\mathbf{x}_\star$ and some estimate $\hat{\boldsymbol{\theta}}$ of the parameter, how do you make a prediction $\hat{y}_\star$?

(C) Now suppose that the parameter $\hat{\boldsymbol{\theta}}$ of the Poisson GLM is estimated using ℓ_2-regularized maximum likelihood estimation. If the test point $\mathbf{x}_\star$ is *orthogonal* to the subspace generated by the training data, what is the distribution $\hat{y}_\star|\mathbf{x}_\star$ predicted by the Poisson GLM model? Prove your answer.

(D) From your answer to part (C), motivate ℓ_1-regularization when the number of features, D, is larger than the number of examples, N.

5 Distances and Optimization Directions

Consider two pairs of distributions with mean and variance parameterization:

Pair 1: Normal(0, 0.0001), Normal(0.1, 0.0001)

Pair 2: Normal(0, 1000), Normal(0.1, 1000)

(A) Make two plots where each plot shows the pdfs for the distributions in the pair.

(B) Compute the Euclidean distance between the parameter vector (mean, variance) for both pairs of distributions. For the same pairs of distributions compute the KL-divergence. Which distance fits intuition better and why?

(C) Assume $\boldsymbol{\theta}_t$ is a parameter for a probability distribution and ρ_t is a scalar. What is the solution to the following optimization algorithm?

$$\max_{\boldsymbol{\theta}_{t+1}} \sum_{i=1}^{n} \log p_{\boldsymbol{\theta}_t}(y_i|\mathbf{x}_i) + (\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t)^\top \left[\nabla_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(y_i|\mathbf{x}_i) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t} \right] - \frac{1}{2\rho_t} ||\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t||_2^2$$

(D) What algorithm does the previous solution correspond to? Does part (B) say anything about why this algorithm might be suboptimal? How would you fix it?