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 $y \in \mathbb{N} = \{0, 1, 2, 3, \ldots\}$, 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?