CS 189/289A Introduction to Machine Learning
Fall 2023 Jennifer Listgarten, Jitendra Malik

DIS3

1 Logistic Regression

Assume that we have n i.i.d. data points $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, where each y_i is a binary label in $\{0, 1\}$. We model the posterior probability as a Bernoulli distribution and the probability for each class is the sigmoid function, i.e., $p(y|\mathbf{x}; \mathbf{w}) = q^y(1-q)^{1-y}$, where $q = s(\mathbf{w}^\top \mathbf{x})$ and $s(\zeta) = \frac{1}{1+e^{-\zeta}}$ is the sigmoid function.

(a) Write out the likelihood and log likelihood functions.

(b) Show that finding maximum likelihood estimate of $\mathbf{w}$ is equivalent to the following optimization problem:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \left[\sum_{i=1}^{n} (1 - y_i)\mathbf{w}^\top \mathbf{x}_i + \log\Big(1 + \exp\Big\{-\mathbf{w}^\top \mathbf{x}_i\Big\}\Big) \right]$$

(c) Comment on whether it is possible to find a closed form maximum likelihood estimate of $\mathbf{w}$, and describe an alternate approach.

2 Gaussian Classification

Let $P(x \mid \omega_i) \sim \mathcal{N}(\mu_i, \sigma^2)$ for a two-category, one-dimensional classification problem with classes ω_1 and ω_2, $P(\omega_1) = P(\omega_2) = 1/2$, and $\mu_2 > \mu_1$.

(a) Find the optimal decision boundary and the corresponding decision rule.

(b) The probability of misclassification (error rate) is:

$$P_e = P((\text{misclassified as } \omega_1) \mid \omega_2)\, P(\omega_2) + P((\text{misclassified as } \omega_2) \mid \omega_1)\, P(\omega_1).$$

Show that the probability of misclassification (error rate) associated with this decision rule is

$$P_e = \frac{1}{\sqrt{2\pi}} \int_a^\infty e^{-z^2/2} dz,$$

where $a = \dfrac{\mu_2 - \mu_1}{2\sigma}$.

(c) What is the limit of P_e as σ goes to 0?

3 Overview of test sets, validation, and cross-validation

In this part, we discuss several issues having to do with test sets and the notions of validation and cross-validation. Open this notebook in datahub and discuss the questions it contains.