Midterm (Group B)

Introduction to Machine Learning
Fall 2024
Instructor: Anna Choromanska

Problem 1 (20 points)

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. samples from Poisson distribution, i.e.

$$f(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \ldots \text{ and } \lambda > 0.$$

Find the maximum likelihood estimator for λ (15 points). Calculate an estimate using this estimator when $x_1 = 1$, $x_2 = 2$, $x_3 = 4$, $x_4 = 2$ (5 points).

Problem 2 (15 points)

Consider 2d family of classifiers given by an origin-centered circles $f(x) = sign(ax^\top x + b)$. What is the VC dimension of this family? Prove it.

Problem 3 (20 points)

Suppose you have a 2-class classification problem, where each class is Gaussian. Let $\theta = \{\alpha, \mu_1, \Sigma_1, \mu_2, \Sigma_2\}$ denote the set of model parameters. Suppose the class probability $p(y|\theta)$ is modelled via the Bernoulli distribution, i.e. $p(y|\theta) = \alpha^y(1-\alpha)^{1-y}$, and the probability of the data $p(x|y,\theta)$ is modelled as $p(x|y,\theta) = \mathcal{N}(x|\mu_y, \Sigma_y)$. Recover the parameters of the model from maximum likelihood approach [2 points for α, 3 points for μ's, and 5 points for Σ's]. Assume the data are i.i.d. and N is the number of data samples. Show all derivations. Next, suppose you want to make a classification decision by assigning proper label y to a given data point x. You decide the label based on Bayes optimal decision $y = \arg\max_{\hat{y}=\{0,1\}} p(\hat{y}|x)$. Prove that the decision boundary is linear when covariances Σ_1 and Σ_2 are equal [10 points].

Problem 4 (10 points)

Prove that a weighted sum of two kernels is a valid kernel: $k(x, y) = \alpha k_1(x, y) + \beta k_2(x, y)$ for $\alpha, \beta \geq 0$, where k_1, k_2 are valid kernels.

Problem 5 (15 points)

Consider the following plot, where we fit the polynomial of order M ($f(x; w) = \sum_{j=0}^{M} w_j x^j$) to the dataset, where $w = [w_0 \quad w_1 \quad \ldots \quad w_M]^\top$ denotes the vector of model weights and the dataset is a collection of 2-dimensional points (x, y). The dataset is represented with the blue circles on the figure.

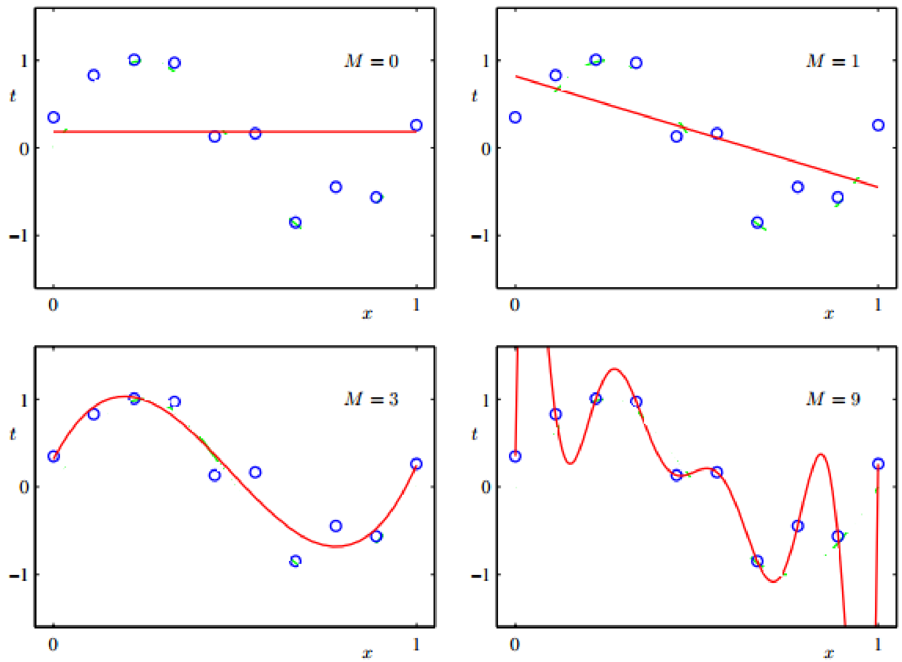

Figure 1: Plots of polynomials having various orders M, shown as red curves, fitted to the data set.

What is the reasonable choice of M and why? Which M correspond to overfitting and which to underfitting and why? (8 points)

Consider any loss function that measures the discrepancy between the target values and the predictions of the model, e.g. squared loss which for a single data point is defined as $L(y_i, f(x_i, w)) = \frac{1}{2}(y_i - f(x_i, w))^2$). Draw a typical behavior of the train and test loss for the optimal setting of model weights as a function of M, where recall that the train loss is the loss computed for a training dataset (the model was trained on this dataset) and the test loss is the loss computed for a test dataset (the model did not see this dataset during training). Indicate overfitting and underfitting regimes (7 points).

Problem 6 (10 points)

Suppose that $w = xy^2$ and $x = r\cos\theta$, $y = r\sin\theta$. Use the chain rule to find $\partial w/\partial r$ when $r = 2$, $\theta = \pi/4$.

Problem 7 (10 points)

Consider 1-dimensional polynomial regression problem when your model

performs label prediction for the i-th example x_i in the training data set using polynomial function:

$$f(x_i; \theta_0, \theta_1, \theta_2, \ldots, \theta_P) = \sum_{p=1}^{P} \theta_p x_i^p + \theta_0.$$

In this case x_i is 1-dimensional. Let y_i denote the true label of the i-th example and let N be the total number of training examples. Parameters of the model $(\theta_0, \theta_1, \theta_2, \ldots, \theta_P)$ are obtained by minimizing the empirical risk given below:

$$R(\theta) = \frac{1}{2N} ||y - X\theta||_2^2 + \theta^T\theta + a^T\theta$$

for some vector a that is given. Write what is y, X, and θ in the formula above. Compute the optimal setting of parameters by setting the gradient of the risk to 0. Explain all steps in your derivations.