Exercises
Advanced Machine Learning
Fall 2019

Series 1. Sep 19th 2019
(Recap IML and important concepts)

Machine Learning Laboratory
Dept. of Computer Science, ETH Zürich
Prof. Joachim M. Buhmann
Web https://ml2.inf.ethz.ch/courses/aml/

Teaching assistant: **Carlos Cotrini**
ccarlos@inf.ethz.ch

Problem 1 (Regression):

1. **Linear Regression** Given is a dataset of inputs $\mathbf{X} \in \mathbb{R}^{n\times d}$ and output variable $\mathbf{y} \in \mathbb{R}^n$. Recall the statistical model for linear regression in homogeneous coordinates is $\hat{y}_i = \sum_{j=1}^{d} x_{ij}\beta_j \implies \hat{\mathbf{y}} = \mathbf{X}\beta$.
   (a) State the residual sum of squares error for this problem in sum and matrix notation.
   (b) Derive the optimal $\hat{\beta}$ by minimizing the loss stated above.
2. **Ridge Regression** In Ridge Regression, an additional term $\lambda\beta^\top\beta$ is added to the residual sum of squares loss.
   (a) Explain the term $\lambda\beta^\top\beta$ in the loss function, what is its impact on the solution? What is the role of the design parameter λ, i.e. what happens for $\lambda \to 0$ and $\lambda \to \infty$?
   (b) Derive the optimal $\hat{\beta}$ by minimizing the ridge loss.

Problem 2 (Comprehension Questions):

1. **Overfitting** A dataset is split into training and validation set. Given are the training and the validation loss as a function of the training time for three neural networks. For every graph, ...
   (a) ... state whether the network is underfitting, reasonable or overfitting and why.
   (b) ... give an example of what you could do to improve the solution.

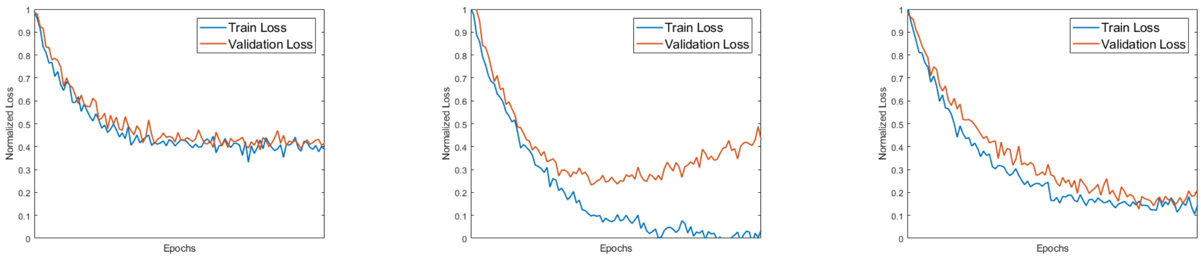

Figure 1: Loss diagrams for networks i. (left), ii. (middle) and iii. (right).

2. **Cross Validation** We want to train a Ridge Regression model and estimate its prediction error. We therefore split the available dataset into a training, validation and test set.
   (a) Explain what the different sets are used for and why the distinction is important.

(b) You choose to use K-fold cross validation to estimate the prediction error for different values of λ. Briefly describe the procedure and explain its advantage over a single split of the taining/validation set.

(c) Is there a systematic tendency affecting the obtained error estimate? Discuss the impact of the number K. How would you choose K in general? What happens for $K \to 1, K \to n$?

3. **Generative vs. Discriminative Modeling** Contrast the generative and the discriminative modeling approach.

   (a) Which probability distributions are they trying to estimate? Which one is harder?

   (b) How is the decision boundary derived in each approach? How does this affect outliers?

   (c) What is the influence of the model specification and available data on the prediction error? Which approach would you prefer? What about unlabeled data?

Problem 3 (EM-Algorithm for Gaussian Mixtures):

Given is a dataset $\mathcal{X} = \{x_i\}_{i=1,\ldots,n}, x_i \in \mathbb{R}^d$, that can be modeled as a mixture of k Gaussian components with parameters $\theta = \{\pi_c, \mu_c, \Sigma_c\}_{c=1,\ldots,k}$, where π_c is the prior probability that a sample is generated by mixture component c, $\mu_c \in \mathbb{R}^d$ and $\Sigma_c \in \mathbb{R}^{d\times d}$ are the mean and covariance matrix of component c. A local optimum for the parameters θ as well as assignment probabilities $P(c|x_i, \theta) = \gamma_{ic}$ can be computed using the EM algorithm.

We are now going to derive the E- and M-Step update equations that maximize the likelihood of our model.

1. Find and expression for the probability of datum x_i in our model, i.e. $P(x_i|\theta)$.
2. Compute the log likelihood of the dataset $L(\mathcal{X}|\theta) = \log(P(\mathcal{X}|\theta))$.
3. We now introduce the binary latent variable $M_{ic} \in \{0, 1\}$, where $M_{ic} = 1$ indicates that x_i is generated by component c, $M_{ic} = 0$ indicates that x_i is not generated by component c. Find an expression for the joint likelihood $P(\mathcal{X}, M|\theta)$ and $L(\mathcal{X}, M|\theta)$.
4. **E-step** During the E-step, we fix our current estimate of θ and call it θ^o. We then calculate the expected assignments $\gamma_{ic} := \mathbb{E}_{M|\mathcal{X},\theta^o}[M_{ic}]$. Calculate γ_{ic} as an expression of π_c and $P(x_i|c, \mu_c, \Sigma_c)$.
5. **M-step** During the M-step, the assignments γ_{ic} are fixed and we optimize $\theta \in \arg\max_\theta Q(\theta)$, where

$$Q(\theta) = \mathbb{E}_{M|\mathcal{X},\theta^o}[L(\mathcal{X}, M|\theta)] = \sum_M P(M \mid \mathcal{X}, \theta^o)L(\mathcal{X}, M \mid \theta).$$

   (a) Compute $Q(\theta)$ as an expression of γ_{ic}, π_c, and $P(x_i|c, \mu_c, \Sigma_c)$.

   (b) Compute the new estimate of μ_c.

   (c) (Hard) Compute the new estimate of Σ_c.

   Hint: Do not compute $\frac{\partial Q}{\partial \Sigma_c}$. Compute instead $\frac{\partial Q}{\partial \Sigma_c^{-1}}$. The following equalities are useful:

$$\frac{\partial}{\partial \Sigma_c^{-1}} \log \left|\Sigma_c^{-1}\right| = \Sigma_c. \qquad \frac{\partial}{\partial \Sigma_c^{-1}} a^\top \Sigma_c^{-1} a = aa^\top, \text{ for } a \in \mathbb{R}^d.$$

   (d) Compute the new estimate of π_c s.t. $\sum_{c=1}^k \pi_c = 1$.

   Hint: The constrained optimization problem $\min_x f(x)$ s.t. $g(x) = 0$ can be solved by introducing the Lagrange multiplier λ: $\mathcal{L} = f(x) + \lambda g(x)$. In this problem, use $\frac{d}{dx}\mathcal{L} = 0$, then eliminate λ using the constraint $\sum_{c=1}^k \pi_c = 1$ to arrive at the optimizer.