582744 Advanced Course in Machine Learning

Exercise 5 Due April 26, 10:00 AM

Rules:

1. Your submission is composed of two parts: (a) A single PDF file containing your answers to all questions, including both the pen and paper questions and the written answers and various plots for the programming questions. (b) a single compressed file (zip or tar.gz) containing your code (and nothing else). If your code is in a single file, it can be sent also a plain source code.
2. The submission should be sent directly to BOTH Arto (`arto.klami@cs.helsinki.fi`) and Aditya (`aditya.jitta@cs.helsinki.fi`).
3. All material must be renamed to your student ID. Always mention your name and student ID also in the written report.
4. The subject field of your email should be [AML][your student ID][exercise 5].
5. Please typeset your work using appropriate software such as LaTeX. However, there is no need to typeset the pen and paper answers – you can also include a scanned hand-written version.
6. The pen and paper exercises can alternatively be returned in paper form during the Tuesday lecture.

This set of exercises is due on Tuesday April 26th, before 10:00 AM.

1 Regularized linear regression (6 pts)

During the lectures it was stated that ridge regression pulls the solution vector towards zero, but we did not clearly state how exactly it happens. If we store individual samples as rows of the input matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$, then the solution for the ridge regression problem

$$\operatorname*{argmin}_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \lambda \|\mathbf{w}\|^2$$

is

$$\mathbf{w}_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

and the solution for the unregularized problem is obtained when $\lambda = 0$. Characterize the relationship between $\mathbf{w}_\lambda$ and $\mathbf{w}_0$, the solution for the unregularized problem. You can either try directly expressing $\mathbf{w}_\lambda$ as a function of $\mathbf{w}_0$, or you can look at the the gradients of the regularized problem at $\mathbf{w}_0$. Describe how the estimate changes for increasing regularization constant λ.

Hint: In case you are unable to solve the problem for vector-valued $\mathbf{x}$, you can start with scalars x instead to get a rough idea of the solution.

2 Gaussian generative classifier (6 pts)

Assume that one-dimensional continuous random variable x and a binary class variable c follow the joint density $p(x, c)$ given by

$$\begin{aligned} p(x|c = 0) &= N(x|0, 1) \\ p(x|c = 1) &= N(x|1, 10^2) \\ p(c = 0) &= 0.4 \\ p(c = 1) &= 0.6 \end{aligned}$$

Here the second parameter of the normal distribution denotes the variance, not standard deviation.

Write down the conditional density for $p(c|x)$ and derive a decision rule for classifying new samples x into the two classes, to create a generative classifier based on this model. What is the expected error for such a classifier? Can you think of algorithms or models that would reach smaller error?

Note that in real uses the parameters of the model would naturally need to be estimated from the data, instead of being known in advance like here.

Hint: Sketching the conditional densities $p(x|c)$ before solving the problem might help.

3 Radial basis functions (6 pts)

Non-linear regression models can be obtained by replacing $\mathbf{w}^T\mathbf{x}$ in a linear model by $\mathbf{w}^T\phi(\mathbf{x})$, where $\phi(\cdot)$ is a feature map, some vector-valued non-linear function of the inputs. Radial basis function (RBF) networks use the feature map $\phi(\mathbf{x}) = [e^{-\|\mathbf{x}-\boldsymbol{\mu}_1\|^2/(2\sigma^2)}, e^{-\|\mathbf{x}-\boldsymbol{\mu}_2\|^2/(2\sigma^2)}, \ldots, e^{-\|\mathbf{x}-\boldsymbol{\mu}_K\|^2/(2\sigma^2)}]$ for some set of K centroids. What kind of information does this feature repsentation capture?

Now imagine we use a collection of N training samples as the centroids, setting $\boldsymbol{\mu}_k = \mathbf{x}_k$, and we use unregularized linear regression to minimize the loss

$$\sum_n (y_n - \mathbf{w}^T\phi(\mathbf{x}_n))^2.$$

Write down the loss function in terms of the original samples and derive the solution that minimizes this loss. Write down also the mean prediction for validation samples $\mathbf{x}$. What happens for the training and validation losses when $\sigma^2 \to 0$ and $\sigma^2 \to \infty$?

The model looks a bit like kernelized (ridge) regression with Gaussian kernels $k_{ij} = e^{-\|\mathbf{x}_i-\mathbf{x}_j\|^2/(2\sigma^2)}$. Is it equivalent? If not, can you explain the difference?

4 SMO algorithm for SVM (6 pts)

The SVM objective is given by

$$\begin{aligned} &\min \sum_{i}^{N} \alpha_i - \frac{1}{2} \sum_{i,j}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq C \quad \forall i \in [1, ..., N] \\ & \sum_{i}^{N} \alpha_i y_i = 0, \end{aligned}$$

where C is some regularization parameter.

The SMO algorithm for optimizing it picks two dual variables α_i and α_j at a time and uses the constraint $\sum_i \alpha_i y_i = 0$ to re-write $\alpha_i = (\eta - \alpha_j y_j) y_i$, where $\eta = -\sum_{k \neq i,j} \alpha_k y_k$. Verify that this expression is correct.

The algorithm then minimizes the loss over α_j. Show that this optimization problem is quadratic. Derive the update rule for determining the optimal value for α_j.