Department of Computer Science

WBI - Machine Learning
Dr. Patrick Jähnichen

Basics of Machine Learning, Winter Term 2017/2018

Exercise sheet 5

Discussion on December 12st and December 15th, 2017

Task 1: Theory

Construct the dual problem for ridge regression. The primal optimization problem is given by

$$\min_{\theta} f_0(\theta)$$

with $f_0(\theta) = (Y - X\theta)^\top (Y - X\theta) + \delta\theta^\top\theta$ and its solution is

$$\theta^*_{\text{primal}} = (X^\top X + \delta\mathbb{I})^{-1} X^\top Y.$$

(a) Reformulate the optimization problem into a constrained one. Use an equality constraint of the form $R = Y - X\theta$.

(b) Form the Lagrangian by incorporating the equality constraint using Lagrange multipliers into a new objective.

(c) Solve the dual problem

$$\max_{\alpha} \min_{\theta, R} \mathcal{L}(\theta, R, \alpha)$$

by first deriving optimal parameters (θ^*, R^*) for the inner minimization.

(d) Plug your results into the Lagrangian to find a function $g(\alpha) = \mathcal{L}(\theta^*, R^*, \alpha)$ and solve the outer problem. (Remember that $\max_x f(x) = \min_x -f(x)$.)

(e) Compare the solutions of the primal and dual problem. When would you choose one over the other? (Note: Plug the optimal α^* as computed in the last step into θ^* from step (c).)

Task 2: Practical

Compare different ridge regression algorithms. In general, the ridge regression cost function is given by

$$J(\theta) = (Y - \Phi(X)\theta)^\top (Y - \Phi(X)\theta) + \delta\theta^\top\theta.$$

Interpreted as an ordinary optimization problem, we can consider different settings for $\Phi(\cdot)$.

- If $\Phi(\cdot)$ is the identity function. we obtain simple linear regression.
- If $\Phi(\cdot)$ transforms single X_i into a vector of polynomials of X_i, we obtain polynomial regression.
- If $\Phi(\cdot)$ defines a set of radial basis functions that transform the input X_i and the number of bases is equal to the number of examples, we obtain kernel regression.

(a) Generate a simple training data set, e.g. take the sine wave and add zero-mean unit-variance Gaussian noise to it, i.e. generate according to $y_i = \sin(x_i) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1)$. Important is, that the underlying function is non-linear.

(b) Fit a polynomial regression to the data set and visualize its predictions. Determine the degree by cross-validation.

(c) Fit a kernel regression to the data. Use radial basis functions of the form

$$\kappa(X_i, \mu_j, \lambda) = \exp\{\frac{1}{\lambda}||X_i - \mu_j||^2\},$$

for each X_i consider basis functions $\{\kappa(X_i, \mu_1, \lambda), \dots, \kappa(X_i, \mu_n, \lambda)\}$ with means $\mu_j = X_j,\ j = \{1, \dots, n\}$. Determine λ using cross-validation and visualize your results.