

SC4/SM8 Advanced Topics in Statistical Machine Learning

Problem Sheet 1

1. This question pertains to the DP-means algorithm.

(a) Show that each iteration of Step 2 locally minimises the following objective:

$$W_\lambda(\{C_k\}, \{\mu_k\}, K) = \sum_{k=1}^K \sum_{i \in C_k} \|x_i - \mu_k\|_2^2 + \lambda K. \quad (1)$$

Answer: For each data item x_i , if its nearest centroid is closer than $\sqrt{\lambda}$, assigning x_i to the cluster will not increase the objective. On the other hand if the nearest centroid is further than $\sqrt{\lambda}$, we can create another cluster centred at x_i , and pay a penalty λ while still decreasing the overall objective.

(b) Conclude that the DP-means algorithm will terminate after a finite number of iterations.

Answer: The reason is the same for K -means. The objective is non-negative, so lower bounded. Each iteration never increase the objective. There are a finite number of partitions, and the algorithm terminates if the partition does not change.

(c) Describe how the tuning parameter λ controls the number of clusters returned by the algorithm.

Answer: Clusters are only created if an item is at least $\sqrt{\lambda}$ away from all existing clusters. So larger values of λ means that clusters are created less frequently, and each cluster will be larger (both in terms of number of items and its size in Euclidean distance). In fact a cluster cannot have a radius larger than $\sqrt{\lambda}$.

2. Show that the second PC is the eigenvector corresponding to the second largest eigenvalue.

Answer: The second PC maximizes the sample variance $\widehat{\text{Var}}(Z^{(2)}) = v_2^\top \widehat{\text{Cov}}(X) v_2$ of the second derived variable among the directions orthogonal to v_1 , that is, it is given by the following optimisation problem:

$$\begin{aligned} \max_{v_2} \quad & v_2^\top S v_2 \\ \text{subject to:} \quad & v_2^\top v_2 = 1, \quad v_1^\top v_2 = 0. \end{aligned}$$

Lagrangian is

$$\mathcal{L}(v_2, \lambda_2, \gamma_2) = v_2^\top S v_2 - \lambda_2 (v_2^\top v_2 - 1) - \gamma_2 v_1^\top v_2$$

and setting the corresponding vector of partial derivatives to zero

$$\frac{\partial \mathcal{L}(v_2, \lambda_2, \gamma_2)}{\partial v_2} = 2Sv_2 - 2\lambda_2 v_2 - \gamma_2 v_1 = 0.$$

Left-multiplying the above by v_1^\top gives $2v_1^\top S v_2 = \gamma_2$. However, since S is symmetric and v_1 is an eigenvector, we have

$$\gamma_2 = 2v_1^\top S v_2 = 2v_2^\top S v_1 = 2\lambda_1 v_2^\top v_1 = 0. \quad (2)$$

Hence $Sv_2 = \lambda_2 v_2$ and similarly as before v_2 must be the eigenvector corresponding to the second largest eigenvalue λ_2 of S .

3. For a given loss function L , the risk R of real-valued $f : \mathcal{X} \rightarrow \mathbb{R}$ is given by the expected loss

$$R(f) = \mathbb{E} [L(Y, f(X))].$$

Derive the optimal regression functions (which minimize the true risk) for the following losses:

- (a) The squared error loss

$$L(Y, f(X)) = (Y - f(X))^2$$

Answer: We have

$$\begin{aligned} R(f) &= \mathbb{E} [(Y - f(X))^2] \\ &= \int \mathbb{E} [(Y - f(X))^2 | X = x] g_X(x) dx, \end{aligned}$$

where g_X is density of X . Thus, it suffices to for every x , minimize:

$$\begin{aligned} &\mathbb{E} [(Y - f(X))^2 | X = x] \\ &= \mathbb{E} [Y^2 | X = x] - 2f(x) \mathbb{E} [Y | X = x] + f(x)^2 \\ &= \text{Var} [Y | X = x] + (\mathbb{E} [Y | X = x] - f(x))^2. \end{aligned}$$

This is clearly minimized by the conditional mean:

$$f(x) = \mathbb{E} [Y | X = x].$$

- (b) The τ -pinball loss, for general $\tau \in (0, 1)$, given by

$$L(Y, f(X)) = 2 \max \{ \tau(Y - f(X)), (\tau - 1)(Y - f(X)) \}.$$

What happens in the case $\tau = 1/2$?

Answer: We want to find $f(x)$ to minimize

$$\mathbb{E} [L(Y, f(X)) | X = x].$$

Note that we can write

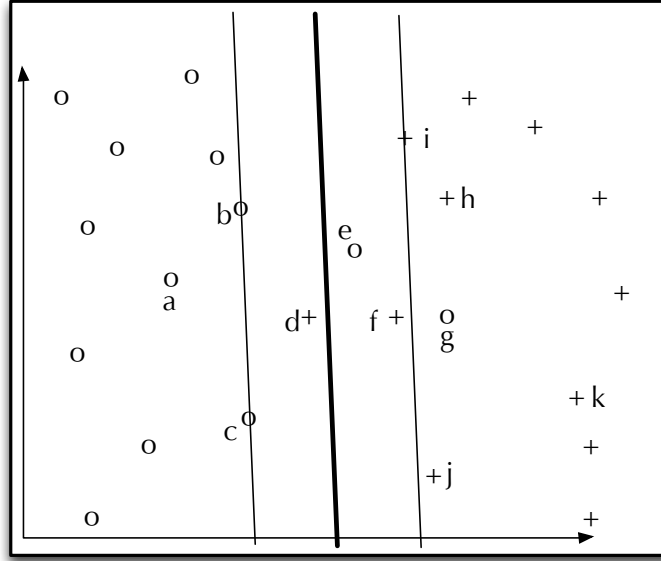
$$L(Y, f(X)) = \begin{cases} 2\tau(Y - f(X)) & \text{if } Y > f(X), \\ 2(\tau - 1)(Y - f(X)) & \text{if } Y \leq f(X). \end{cases}$$

Differentiating with respect to $f(x)$ and setting to zero, we obtain

$$2\tau \mathbb{P}(Y > f(x) | X = x) + 2(\tau - 1) \mathbb{P}(Y \leq f(x) | X = x) = 0,$$

leading to $\mathbb{P}(Y \leq f(x) | X = x) = \tau$ so the optimal $f(x)$ is the τ -quantile of the conditional distribution function $\mathbb{P}(Y \leq y | X = x)$. In the special case $\tau = 1/2$, we obtain L1 loss and the conditional median as the optimal regressor.

4. The figure below shows a binary classification dataset and the optimal the decision boundary and margins of a soft-margin C -SVM for some value C .



- (a) Which of the points a, \dots, k are definitely support vectors? What can you say about points b, c, i ? Can they be non, margin, or non-margin support vectors?

Answer: The points d, e, f, g are (non-margin) support vectors. b, c, i can in fact be non-SVs (if $\alpha = 0$), margin support vectors (if $0 < \alpha < C$) or non-margin support vectors (if $\alpha = C$).

- (b) For points a, b and d what are the range of possible values for the corresponding dual variables?

Answer: Point a , the dual variable $\alpha_a = 0$ (removing a does not affect the boundary).
 Point b , the dual variable $\alpha_b \in [0, C]$ (typically this is a margin support vector and typically $\alpha_b \in (0, C)$).
 Point d , the dual variable $\alpha_d = C$ (margin penalty).

5. Parameter C in C -SVM can sometimes be hard to interpret. An alternative parametrization is given by ν -SVM:

$$\min_{w, b, \rho, \xi} \left(\frac{1}{2} \|w\|^2 - \nu \rho + \frac{1}{n} \sum_{i=1}^n \xi_i \right)$$

subject to

$$\begin{aligned} \rho &\geq 0, \\ \xi_i &\geq 0, \\ y_i (w^\top x_i + b) &\geq \rho - \xi_i. \end{aligned}$$

(note that we now directly adjust the constraint threshold ρ).

Using complementary slackness, show that ν is an upper bound on the proportion of non-margin support vectors (margin errors) and a lower bound on the proportion of all support vectors with non-zero weight (both those on the margin and margin errors). You can assume that $\rho > 0$ at the optimum (non-zero margin).

Answer: The Lagrangian is given by

$$\frac{1}{2}\|w\|^2 - \nu\rho + \frac{1}{n} \sum_{i=1}^n \xi_i + \sum_{i=1}^n \alpha_i \left(\rho - y_i(w^\top x_i + b) - \xi_i \right) + \sum_{i=1}^n \beta_i (-\xi_i) + \gamma(-\rho),$$

for $\alpha_i \geq 0$, $\beta_i \geq 0$, $\gamma \geq 0$. Differentiating w.r.t. to the primal variables w , b , ξ , ρ and setting to zero gives

$$\begin{aligned} w &= \sum_{i=1}^n \alpha_i y_i x_i, \\ \sum_{i=1}^n \alpha_i y_i &= 0, \\ \alpha_i + \beta_i &= \frac{1}{n}, \\ \nu &= \sum_{i=1}^n \alpha_i - \gamma. \end{aligned}$$

Thus $\alpha_i \in [0, \frac{1}{n}]$, and $\nu \leq \sum_{i=1}^n \alpha_i$. Assume $\rho > 0$ at the global solution. This just means that the margin is non-zero. Then, by complementary slackness $\gamma = 0$, and $\sum_{i=1}^n \alpha_i = \nu$.

- Non-margin support vectors are those with $\alpha_i = \frac{1}{n}$, for which by complementary slackness $y_i(w^\top x_i + b) = \rho - \xi_i$, and from $\alpha_i + \beta_i = \frac{1}{n}$, $\beta_i = 0$, so potentially $\xi_i > 0$ (margin error). Denote this set by $N(\alpha)$. Then

$$\frac{|N(\alpha)|}{n} = \sum_{i \in N(\alpha)} \frac{1}{n} = \sum_{i \in N(\alpha)} \alpha_i \leq \sum_{i=1}^n \alpha_i = \nu.$$

- Denote margin support vectors, i.e. those with $\alpha_i \in (0, \frac{1}{n})$, with $M(\alpha)$. For these $\beta_i > 0$ and thus $\xi_i = 0$, so $y_i(w^\top x_i + b) = \rho$. Furthermore,

$$\nu = \sum_{i=1}^n \alpha_i = \sum_{i \in N(\alpha)} \frac{1}{n} + \sum_{i \in M(\alpha)} \alpha_i \leq \sum_{i \in M(\alpha) \cup N(\alpha)} \frac{1}{n} = \frac{|N(\alpha) + M(\alpha)|}{n}.$$

Thus ν is an upper bound on the number of margin errors and a lower bound on the number of support vectors.

6. Consider the regression problem to the real-valued output $y \in \mathbb{R}$. Let $\epsilon > 0$ and define the ϵ -insensitive loss function L_ϵ as

$$L_\epsilon(y, f(x)) = \begin{cases} 0 & \text{if } |y - f(x)| < \epsilon, \\ |y - f(x)| - \epsilon & \text{otherwise,} \end{cases}$$

and the regularized empirical risk objective defined as

$$J(w, b) = C \sum_{i=1}^n L_\epsilon(y_i, f(x_i)) + \frac{1}{2}\|w\|_2^2,$$

where we used a linear model $f(x) = w^\top x + b$ for regression functions.

- (a) Introduce the slack variables $\xi_i^+ = \max\{y_i - f(x_i) - \epsilon, 0\}$ and $\xi_i^- = \max\{f(x_i) - y_i - \epsilon, 0\}$. Verify that $L_\epsilon(y_i, f(x_i)) = \xi_i^+ + \xi_i^-$.

Answer: $L_\epsilon(y_i, f(x_i))$ is non-zero only if either $y_i - f(x_i) > \epsilon$, in which case it is equal to ξ_i^+ , or if $f(x_i) - y_i > \epsilon$, in which case it is equal to ξ_i^- . Furthermore, if $\xi_i^+ > 0$, then $f(x_i) - y_i - \epsilon < -2\epsilon$, so $\xi_i^- = 0$.

- (b) Re-express the regularized empirical risk objective $J(w, b)$ as a constrained optimization problem over w, b, ξ^+ and ξ^- . Write down Lagrangian and show that the dual problem can be written as

$$\max_{\alpha^+, \alpha^-} \left\{ -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n (\alpha_i^+ - \alpha_i^-)(\alpha_j^+ - \alpha_j^-) x_i^\top x_j + \sum_{i=1}^n (\alpha_i^+ - \alpha_i^-) y_i - \epsilon \sum_{i=1}^n (\alpha_i^+ + \alpha_i^-) \right\},$$

subject to

$$\sum_{i=1}^n (\alpha_i^+ - \alpha_i^-) = 0, \quad \alpha_i^+ \in [0, C], \quad \alpha_i^- \in [0, C], \quad i = 1, \dots, n.$$

Answer: The primal problem is given by

$$\min_{w, b, \xi^+, \xi^-} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n (\xi_i^+ + \xi_i^-)$$

subject to

$$\begin{aligned} \xi_i^+ &\geq 0 & \xi_i^+ &\geq y_i - f(x_i) - \epsilon, i = 1, \dots, n, \\ \xi_i^- &\geq 0 & \xi_i^- &\geq f(x_i) - y_i - \epsilon, i = 1, \dots, n. \end{aligned}$$

Lagrangian is given by

$$\begin{aligned} \mathcal{L} = & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n (\xi_i^+ + \xi_i^-) - \sum_{i=1}^n (\nu_i^+ \xi_i^+ + \nu_i^- \xi_i^-) \\ & + \sum_{i=1}^n \alpha_i^+ (y_i - f(x_i) - \epsilon - \xi_i^+) + \sum_{i=1}^n \alpha_i^- (f(x_i) - y_i - \epsilon - \xi_i^-), \end{aligned}$$

where the Lagrange multipliers $\alpha_i^+, \alpha_i^-, \nu_i^+, \nu_i^-$ are all ≥ 0 . By differentiating with respect to primal variables and setting to zero,

$$\begin{aligned} \partial_b \mathcal{L} &= \sum_{i=1}^n (\alpha_i^+ - \alpha_i^-) = 0, \\ \partial_w \mathcal{L} &= w - \sum_{i=1}^n (\alpha_i^+ - \alpha_i^-) x_i = 0, \\ \partial_{\xi_i^\pm} \mathcal{L} &= C - \alpha_i^\pm - \nu_i^\pm = 0, \end{aligned}$$

and substituting back in, we obtain the desired dual problem.

- (c) Considering derivatives of the Lagrangian and complementary slackness, express the weight vector w using dual coefficients α_i^+ and α_i^- . Show that those examples (x_i, y_i) which lie outside of the ϵ -insensitive tube around f , must have corresponding $\alpha_i^+ = C$ or $\alpha_i^- = C$ and that those examples (x_i, y_i) for which $|f(x_i) - y_i| < \epsilon$ (they lie strictly inside the ϵ -tube), must have $\alpha_i^+ = \alpha_i^- = 0$. How can you compute b using the dual solution?

Answer: The partial derivatives of the Lagrangian with respect to primal variables have to vanish for optimality. From $\partial_w \mathcal{L}$, it follows that $w = \sum_{i=1}^n (\alpha_i^+ - \alpha_i^-) x_i$. Also, from $\partial_{\xi_i^\pm} \mathcal{L}$, it follows that $\nu_i^\pm = C - \alpha_i^\pm$. Now, by complementary slackness, also $\nu_i^\pm \xi_i^\pm = 0$ for all i . Hence, if an error larger than ϵ is committed ($\xi_i^\pm > 0$), it must be that $\nu_i^\pm = 0$ and hence $\alpha_i^\pm = C$. On the other hand, with $|f(x_i) - y_i| < \epsilon$, we have $\xi_i^+ = \xi_i^- = 0$ and $f(x_i) - y_i - \epsilon$ and $y_i - f(x_i) - \epsilon$ are both nonzero, so $\alpha_i^+ = \alpha_i^- = 0$ by complementary slackness. Therefore, we have a sparse expansion of w in terms of x_i . To compute b , take any $\alpha_i^\pm \in (0, C)$ – it must be $|y_i - f(x_i)| = \epsilon$, so we can compute $b = y_i - w^\top x_i \pm \epsilon$.

7. **(Kernel Ridge Regression)** Let $(x_i, y_i)_{i=1}^n$ be our dataset, with $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$. Classical linear regression can be formulated as empirical risk minimization, where the model is to predict y using a class of functions $f(x) = w^\top x$, parametrized by vector $w \in \mathbb{R}^p$ using the squared loss, i.e. we minimize

$$\hat{R}(w) = \frac{1}{n} \sum_{i=1}^n (y_i - w^\top x_i)^2.$$

- (a) Show that the optimal parameter vector is

$$\hat{w} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

where \mathbf{X} is a $n \times p$ matrix with i th row given by x_i^\top , and \mathbf{y} is a $n \times 1$ column vector with i -th entry y_i .

Answer: We can write the empirical risk as

$$\frac{1}{n} \|\mathbf{y} - \mathbf{X}w\|_2^2$$

Differentiating wrt w and setting to 0,

$$\begin{aligned} (\mathbf{X}w - \mathbf{y})^\top \mathbf{X} &= 0 \\ w^\top (\mathbf{X}^\top \mathbf{X}) - \mathbf{y}^\top \mathbf{X} &= 0 \\ \hat{w} &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \end{aligned}$$

- (b) Consider regularizing our empirical risk by incorporating an L_2 regularizer. That is, find w minimizing

$$\frac{1}{n} \sum_{i=1}^n (y_i - w^\top x_i)^2 + \frac{\lambda}{n} \|w\|_2^2$$

Show that the optimal parameter is given by the ridge regression estimator

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

Answer: The objective becomes:

$$\frac{1}{n} \|\mathbf{y} - \mathbf{X}w\|_2^2 + \frac{\lambda}{n} \|w\|_2^2$$

Again differentiating and setting derivative to 0,

$$\begin{aligned} (\mathbf{X}w - \mathbf{y})^\top \mathbf{X} + \lambda w^\top &= 0 \\ w^\top (\lambda I + \mathbf{X}^\top \mathbf{X}) - \mathbf{y}^\top \mathbf{X} &= 0 \\ \hat{w} &= (\lambda I + \mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \end{aligned}$$

- (c) Suppose that we now wish to introduce nonlinearities into the model, by transforming $x \mapsto \varphi(x)$. Let Φ be a matrix with i th row given by $\varphi(x_i)^\top$. The optimal parameters \hat{w} would then be given by (previous part):

$$\hat{w} = (\Phi^\top \Phi + \lambda I)^{-1} \Phi^\top \mathbf{y}.$$

Can we make predictions without computing \hat{w} ?

First, express the predicted y values on the training set, $\Phi \hat{w}$, only in terms of \mathbf{y} and the Gram matrix $\mathbf{K} = \Phi \Phi^\top$, with $\mathbf{K}_{ij} = \varphi(x_i)^\top \varphi(x_j) = k(x_i, x_j)$ where k is some kernel function. Then, compute an expression for the value of y_* predicted by the model at an unseen test vector x_* .

Hint: You may find it useful to first prove that:

$$(\Phi^\top \Phi + \lambda I)^{-1} \Phi^\top = \Phi^\top (\Phi \Phi^\top + \lambda I)^{-1}.$$

Answer: To prove the identity in the hint, we have

$$\begin{aligned} (\Phi^\top \Phi + \lambda I)^{-1} \Phi^\top &= (\Phi^\top \Phi + \lambda I)^{-1} \Phi^\top (\Phi \Phi^\top + \lambda I) (\Phi \Phi^\top + \lambda I)^{-1} \\ &= (\Phi^\top \Phi + \lambda I)^{-1} (\Phi^\top \Phi + \lambda I) \Phi^\top (\Phi \Phi^\top + \lambda I)^{-1} \\ &= \Phi^\top (\Phi \Phi^\top + \lambda I)^{-1}. \end{aligned}$$

Now, using Φ instead of \mathbf{X} , we would get

$$\hat{w} = (\lambda I + \Phi^\top \Phi)^{-1} \Phi^\top \mathbf{y}$$

instead. Multiply by Φ ,

$$\begin{aligned} \Phi \hat{w} &= \Phi (\lambda I + \Phi^\top \Phi)^{-1} \Phi^\top \mathbf{y} \\ &= \Phi \Phi^\top (\lambda I + \Phi \Phi^\top)^{-1} \mathbf{y} \\ &= \mathbf{K} (\lambda I + \mathbf{K})^{-1} \mathbf{y}. \end{aligned}$$

Finally, for a test vector x_* , let $\varphi_* = \varphi(x_*)$. Then the prediction is $\varphi_*^\top \hat{w}$, which gives

$$\begin{aligned} & \lambda^{-1}(\varphi_*^\top \Phi^\top - \varphi_*^\top \Phi^\top (\lambda I + \mathbf{K})^{-1} \mathbf{K}) \mathbf{y} \\ &= \lambda^{-1} \varphi_*^\top \Phi^\top (\lambda I + \mathbf{K})^{-1} (\lambda I + \mathbf{K} - \mathbf{K}) \mathbf{y} \\ &= \varphi_*^\top \Phi^\top (\lambda I + \mathbf{K})^{-1} \mathbf{y} \end{aligned}$$

where we note that $\varphi_*^\top \Phi^\top$ is a row vector with i th entry $k(x_*, x_i)$.

In particular, the nonlinear model can be “kernelized” and all computations can be carried out without explicit computation of $\varphi(x)$ nor of the “primal” weight vector \hat{w} .

8. Denote $\sigma(t) = 1/(1 + e^{-t})$. Verify that the ERM corresponding to the logistic loss over the functions of the form $f(x) = w^\top \varphi(x)$ can be written as

$$\min_w \sum_{i=1}^n -\log \sigma(y_i w^\top \varphi(x_i)) + \lambda \|w\|_2^2 \quad (3)$$

and is a convex optimisation problem in w . Assume that you can write $w = \sum_{i=1}^n \alpha_i \varphi(x_i)$. Show that the criterion in (2) is also convex in the so called dual coefficients $\alpha \in \mathbb{R}^n$. [Hint: $\sigma'(t) = \sigma(t)\sigma(-t)$]

Answer:

The first step, albeit without regularisation, was derived in the lecture notes - it just suffices to check that the logistic function $\rho(t) = \log(1 + e^{-t}) = -\log \sigma(t)$ is convex, which is true since $\rho''(t) = e^t/(1 + e^t)^2 \geq 0$. Regularisation just adds λI to the Hessian.

For the second part, we write $w^\top \varphi(x_i) = \alpha^\top \mathbf{k}_i$ where $\mathbf{k}_i = [k(x_i, x_1), \dots, k(x_i, x_n)]^\top$ to get the problem expressed in terms of α :

$$\min_\alpha \sum_{i=1}^n -\log \sigma(y_i \alpha^\top \mathbf{k}_i) + \lambda \alpha^\top \mathbf{K} \alpha, \quad (4)$$

with Hessian

$$\frac{\partial^2 J}{\partial \alpha \partial \alpha^\top} = \sum_{i=1}^n \sigma(y_i \alpha^\top \mathbf{k}_i) \sigma(-y_i \alpha^\top \mathbf{k}_i) \mathbf{k}_i \mathbf{k}_i^\top + 2\lambda \mathbf{K},$$

which is positive semidefinite. There is actually nothing special here about the logistic loss - the same argument works for any differentiable loss convex in $yf(x)$.