Indian Institute of Technology Kanpur
CS771 Introduction to Machine Learning

Instructor: Purushottam Kar
Date: August 27, 2017
Total: 80 marks

ASSIGNMENT 1

Problem 1.1 (V for Voronoi)**.** Recall the learning with prototypes problem. Consider a two class problem where the prototypes are the points $(1,0)$ (green) and $(0,1)$ (red). Calculate the decision boundary when we use the learning with prototypes rule but with the following *Mahalanobis* metrics. In the following, $\mathbf{z}^1, \mathbf{z}^2 \in \mathbb{R}^2$ denote two points on the real plane

1. $d(\mathbf{z}^1, \mathbf{z}^2) = \left\langle \mathbf{z}^1 - \mathbf{z}^2, U(\mathbf{z}^1 - \mathbf{z}^2) \right\rangle$, where $U = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}$
2. $d(\mathbf{z}^1, \mathbf{z}^2) = \left\langle \mathbf{z}^1 - \mathbf{z}^2, V(\mathbf{z}^1 - \mathbf{z}^2) \right\rangle$, where $V = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$

(5+5=10 marks)

Solution. *Note 1*: the expressions in the question actually depict the squared Mahalanobis distances. This was an error but did not affect the solution since for positive $x, y \in \mathbb{R}$, we have $x \geq y$ iff $x^2 \geq y^2$.

Note 2: some texts use the form $\left\langle \mathbf{z}^1 - \mathbf{z}^2, U^{-1}(\mathbf{z}^1 - \mathbf{z}^2) \right\rangle$ to define the Mahalanobis distance (e.g. Wikipedia article on Mahalanobis distance) whereas others (e.g. the LMNN paper itself) use the notation we have used i.e. $\left\langle \mathbf{z}^1 - \mathbf{z}^2, U(\mathbf{z}^1 - \mathbf{z}^2) \right\rangle$.

If $\mathbf{a}, \mathbf{b} \in \mathbb{R}^2$ denote the green and red prototypes respectively, and $\mathbf{z} \in \mathbb{R}^2$ denotes a point on the decision boundary, then $\mathbf{z}$ must satisfy $d(\mathbf{z}, \mathbf{a}) = d(\mathbf{z}, \mathbf{b})$ which translates to

$$\mathbf{z}^\top U(\mathbf{b} - \mathbf{a}) = \frac{\mathbf{b}^\top U \mathbf{b} - \mathbf{a}^\top U \mathbf{a}}{2},$$

which translates to the following decision boundaries in the two cases (with the notation $\mathbf{z} = (x, y)^\top$.

1. Decision boundary $y = 3x - 1$
2. Decision boundary $x = 0.5$

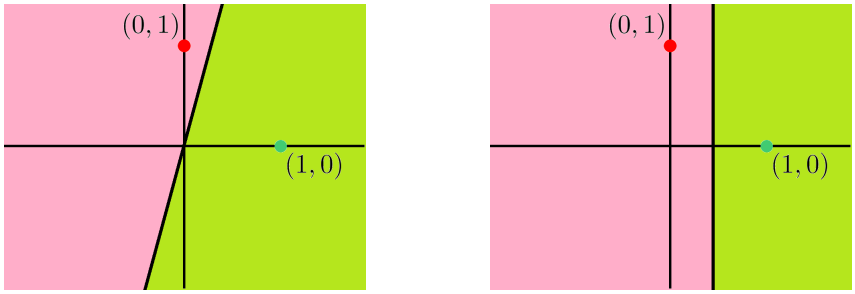

Figure 1: Solutions to parts 1 and 2

Problem 1.2 (PML For Constraints). Consider the following constrained least-squares regression problem on a data set $(\mathbf{x}^i, y^i)_{i=1,\ldots,n}$, where $\mathbf{x}^i \in \mathbb{R}^d$ and $y^i \in \mathbb{R}$.

$$\hat{\mathbf{w}}_{\text{cls}} = \arg\min \sum_{i=1}^{n} (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2$$
$$\text{s.t. } \|\mathbf{w}\|_2 \leq r.$$

Design a likelihood distribution (on the responses, conditioned on the data covariates $\mathbf{x}$) and prior distribution (on the parameter) such that $\hat{\mathbf{w}}_{\text{cls}}$ is the MAP estimate for your model. Give explicit forms for the density functions of your likelihood and prior distributions. The above shows that PML approaches can also lead to constrained optimization problems. (5 marks)

Solution. Likelihood density function: $\mathbb{P}\left[y^i \,|\, \mathbf{x}^i, \mathbf{w}\right] = \mathcal{N}(\langle \mathbf{w}, \mathbf{x}^i \rangle, \sigma)$, for any $\sigma > 0$.
Prior density function: $\mathbb{P}[\mathbf{w}] = 0$ if $\|\mathbf{w}\|_2 > r$, else $\mathbb{P}[\mathbf{w}] = \frac{1}{\text{Vol}(d,r)}$,
where $\text{Vol}(d, r)$ is the volume of the Euclidean ball of radius r in d-dimensions. We have $\text{Vol}(d, r) = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)} \cdot r^d$, where Γ is the Gamma function that extends the factorial function.

Thus, the likelihood is the good old normal/Gaussian likelihood that we have used in ridge regression etc. The prior is more interesting. The prior has zero density outside the ball of radius r i.e. its support is only the ball. Inside the ball, the prior is a uniform prior.

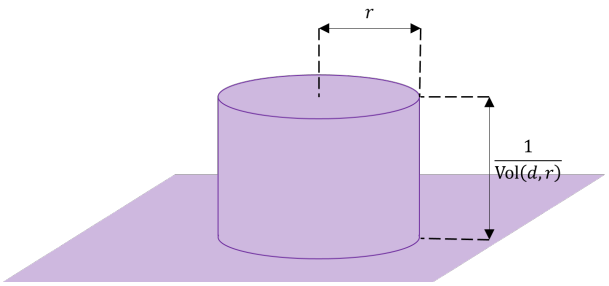

Figure 2: Depiction of the density function of the prior distribution in $d = 2$ dimensions

Note: Some students have tried to exploit the equivalence between regularization and constraints to solve the problem. They took a discussion on Piazza that took a constrained optimization problem with a constraint $\|\mathbf{w}\|_2 \leq r$ and convert it into an unconstrained optimization problem with a regularizer $\lambda \cdot \|\mathbf{w}\|_2^2$ and then used that to construct the prior. This is wrong! Why? Because the value of λ such a technique would get would depend not only on r but also on the data $\mathbf{x}^i, y^i$. Yes, it is true that for every optimization problem of the following kind

$$\arg\min \sum_{i=1}^{n} (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2$$
$$\text{s.t. } \|\mathbf{w}\|_2 \leq r.$$

there exists an optimization problem

$$\arg\min \sum_{i=1}^{n} (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2 + \lambda \cdot \|\mathbf{w}\|_2^2$$

so that the two problems share a solution. However, the λ value for which the two problems are equivalent, changes when one changes the data points $\mathbf{x}^i, y^i$. So any one using this route is essentially proposing a prior that can be defined only *after* data has been seen which defeats the whole purpose of a prior.

Problem 1.3 (Fun with Features). Consider the following *feature-regularized* least-squares regression problem on a data set $(\mathbf{x}^i, y^i)_{i=1,\ldots,n}$, where $\mathbf{x}^i \in \mathbb{R}^d$, $y^i \in \mathbb{R}$, and $\alpha_j > 0$ for $j \in [d]$.

$$\hat{\mathbf{w}}_{\text{fr}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \sum_{i=1}^{n} (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2 + \sum_{j=1}^{d} \alpha_i (\mathbf{w}_i)^2$$

Design a likelihood and prior distribution such that $\hat{\mathbf{w}}_{\text{fr}}$ is the MAP estimate for your model. Give explicit forms for all distributions. It turns out that just as there exists a closed form expression for the solution to the L_2-regularized least-squares problem, one exists for this problem too. Find a closed-form expression for $\hat{\mathbf{w}}_{\text{fr}}$. (5+5=10 marks)

Solution. Prior density function: $\mathbb{P}[\mathbf{w}] = \mathcal{N}(\mathbf{0}, \Sigma)$, for $\Sigma = \text{diag}(\frac{1}{\alpha_1}, \frac{1}{\alpha_2}, \ldots, \frac{1}{\alpha_d})$. The Σ matrix is a diagonal matrix.
Likelihood density function: $\mathbb{P}\left[y^i \,|\, \mathbf{x}^i, \mathbf{w}\right] = \mathcal{N}(\langle \mathbf{w}, \mathbf{x}^i \rangle, \sigma)$, for $\sigma = 1$.
Also valid are solutions that use $\sigma = c$ and $\Sigma = \text{diag}(\frac{c^2}{\alpha_1}, \frac{c^2}{\alpha_2}, \ldots, \frac{c^2}{\alpha_d})$ fo some $c > 0$.
Rewrite the optimization problem as $\min_{\mathbf{w}} \left\|X^\top \mathbf{w} - \mathbf{y}\right\|_2^2 + \mathbf{w}^\top \Sigma^{-1} \mathbf{w}$. Using first order optimality condition we get the solution in closed form as $\hat{\mathbf{w}}_{\text{fr}} = \left(XX^\top + \Sigma^{-1}\right)^{-1} X\mathbf{y}$. Since all $\alpha_i > 0$, the matrix $XX^\top + \Sigma^{-1}$ is invertible.

Problem 1.4 (Break Free from Constraints). Recall the OVA approach to multi-classification. Let us use a dataset $(\mathbf{x}^i, y^i)_{i=1,\ldots,n}$, where $\mathbf{x}^i \in \mathbb{R}^d$ and $y^i \in [K]$ i.e. there are K classes. Denote using $\mathbf{W} = [\mathbf{w}^1, \ldots, \mathbf{w}^K] \in \mathbb{R}^{d\times K}$, the set of K linear models that make up the OVA classifier. The Crammer-Singer formulation $(P1)$ for a single machine learner for multi-classification is

$$\begin{aligned} \left\{\widehat{\mathbf{W}}, \{\hat{\xi}_i\}\right\} = \arg\min_{\mathbf{W}, \{\xi_i\}} & \sum_{k=1}^{K} \left\|\mathbf{w}^k\right\|_2^2 + \sum_{i=1}^{n} \xi_i \\ \text{s.t. } & \left\langle \mathbf{w}^{y^i}, \mathbf{x}^i \right\rangle \geq \left\langle \mathbf{w}^k, \mathbf{x}^i \right\rangle + 1 - \xi_i, \forall i, \forall k \neq y^i \qquad (P1) \\ & \xi_i \geq 0, \text{ for all } i \end{aligned}$$

Show that $(P1)$ is equivalent to the following unconstrained formulation $(P2)$

$$\widehat{\mathbf{W}} = \arg\min_{\mathbf{W}} \sum_{k=1}^{K} \left\|\mathbf{w}^k\right\|_2^2 + \sum_{i=1}^{n} \ell_{\text{cs}}(y^i, \boldsymbol{\eta}^i) \qquad (P2),$$

where $\boldsymbol{\eta}^i = \langle \mathbf{W}, \mathbf{x}^i \rangle$ and

$$\ell_{\text{cs}}(y^i, \boldsymbol{\eta}^i) = [1 + \max_{k\neq y} \boldsymbol{\eta}_k^i - \boldsymbol{\eta}_y^i]_+$$

To show equivalence, you will have to show that if $\left\{\mathbf{W}^0, \{\xi_i^0\}\right\}$ are an optimum for (P_1) then $\mathbf{W}^0$ must be an optimum for $(P2)$, as well as if $\mathbf{W}^1$ is an optimum for $(P2)$ then there must exist $\left\{\xi_i^1\right\} \geq 0$ such that $\left\{\mathbf{W}^1, \{\xi_i^1\}\right\}$ are an optimum for (P_1). (15 marks)

Solution. A very useful result for this analysis is the following:

Lemma 1.1. Suppose we have a model $\mathbf{W} = [\mathbf{w}^1, \ldots, \mathbf{w}^K] \in \mathbb{R}^{d\times K}$ and some real numbers $\xi_i, i \in [n]$ such that $\{\mathbf{W}, \{\xi_i\}\}$ satisfy all the constraints of (P_1). Then we must have $\xi_i \geq \ell_{\text{cs}}(y^i, \langle \mathbf{W}, \mathbf{x}^i \rangle)$.

Proof. For any i, denote $\ell^i := \ell_{\text{cs}}(y^i, \boldsymbol{\eta}^i) = [1 + \max_{k\neq y^i} \boldsymbol{\eta}_k^i - \boldsymbol{\eta}_{y^i}^i]_+$, where $\boldsymbol{\eta}^i = \langle \mathbf{W}, \mathbf{x}^i \rangle$. We want to show $\xi^i \geq \ell^i$ for all i. Consider the following three cases

1. Case 0, $\ell^i < 0$. This can never happen as the positive-part function never takes negative values.

2. Case 1, $\ell^i = 0$. Since $\{\mathbf{W}, \{\xi_i\}\}$ satisfy all the constraints of (P_1), we must have $\xi^i \geq 0$. Thus we must also have $\xi^i \geq \ell^i$.

3. Case 2, $\ell^i > 0$. By definition of the positive-part function $[\cdot]_+$, this corresponds to the case where $1+\max_{k\neq y} \boldsymbol{\eta}^i_k - \boldsymbol{\eta}^i_{y^i} > 0$ which means for some $k \neq y^i$, we must have $1+\boldsymbol{\eta}^i_k - \boldsymbol{\eta}^i_{y^i} > 0$. Let $\tilde{k} = \arg\max_{k\neq y} \boldsymbol{\eta}^i_k$. It is easy to see that $\ell = 1 + \boldsymbol{\eta}^i_{\tilde{k}} - \boldsymbol{\eta}^i_{y^i}$. However, since ξ_i satisfies $\left\langle \mathbf{w}^{y^i}, \mathbf{x}^i \right\rangle \geq \left\langle \mathbf{w}^k, \mathbf{x}^i \right\rangle + 1 - \xi_i$ for all $k \neq y^i$, it must satisfy this for $\tilde{k}$ as well. This gives us $\xi^i \geq \ell^i$.

This concludes the proof. □

Using the above lemma, we will analyse the two directions in two cases

1. Suppose $\mathbf{W} = [\mathbf{w}^1, \ldots, \mathbf{w}^K] \in \mathbb{R}^{d\times K}$ is an optimum for $(P2)$. Let $\xi_i := \ell_{\text{cs}}(y^i, \left\langle \mathbf{W}, \mathbf{x}^i \right\rangle)$. It is easy to see that by the definition of the loss function $\ell_{\text{cs}}(y^i, \boldsymbol{\eta}^i)$, we have $\xi_i \geq 0$ for all i (the positive-part function $[\cdot]_+$ only takes non-negative values) as well as $\left\langle \mathbf{w}^{y^i}, \mathbf{x}^i \right\rangle \geq \left\langle \mathbf{w}^k, \mathbf{x}^i \right\rangle + 1 - \xi_i$ for all i (by an argument similar to the one used to prove the lemma above). Thus, $\{\mathbf{W}, \{\xi_i\}\}$ satisfy the constraints of (P_1). Now suppose that $\{\mathbf{W}, \{\xi_i\}\}$ is not an optimum for (P_1) and there exists a better optimum for (P_1). This means there must exist $\left\{\tilde{\mathbf{W}}, \{\tilde{\xi}_i\}\right\}$ such that they satisfy all the constraints of (P_1) but we have

$$\sum_{k=1}^{K} \left\|\tilde{\mathbf{w}}^k\right\|_2^2 + \sum_{i=1}^{n} \tilde{\xi}_i < \sum_{k=1}^{K} \left\|\mathbf{w}^k\right\|_2^2 + \sum_{i=1}^{n} \xi_i$$

Now, since $\left\{\tilde{\mathbf{W}}, \{\tilde{\xi}_i\}\right\}$ satisfy all the constraints of (P_1), we must have $\tilde{\xi}_i \geq \ell_{\text{cs}}(y^i, \left\langle \tilde{\mathbf{W}}, \mathbf{x}^i \right\rangle)$ (see lemma above). Also notice that we set $\xi_i := \ell_{\text{cs}}(y^i, \left\langle \mathbf{W}, \mathbf{x}^i \right\rangle)$. This gives us

$$\sum_{k=1}^{K} \left\|\tilde{\mathbf{w}}^k\right\|_2^2 + \sum_{i=1}^{n} \ell_{\text{cs}}(y^i, \left\langle \tilde{\mathbf{W}}, \mathbf{x}^i \right\rangle) < \sum_{k=1}^{K} \left\|\mathbf{w}^k\right\|_2^2 + \sum_{i=1}^{n} \ell_{\text{cs}}(y^i, \langle \mathbf{W}, \mathbf{x}^i \rangle),$$

which means that $\tilde{\mathbf{W}}$ is a better solution to (P_2) that $\mathbf{W}$. But this contradicts the (assumed) fact that $\mathbf{W}$ is an optimum for $(P2)$. This means that $\{\mathbf{W}, \{\xi_i\}\}$ must be an optimum for (P_1).

2. Suppose $\{\mathbf{W}, \{\xi_i\}\}$ is an optimum for (P_1) and also suppose that $\mathbf{W}$ is not an optimum for (P_2). This means that there must exist some $\tilde{\mathbf{W}}$ such that

$$\sum_{k=1}^{K} \left\|\tilde{\mathbf{w}}^k\right\|_2^2 + \sum_{i=1}^{n} \ell_{\text{cs}}(y^i, \left\langle \tilde{\mathbf{W}}, \mathbf{x}^i \right\rangle) < \sum_{k=1}^{K} \left\|\mathbf{w}^k\right\|_2^2 + \sum_{i=1}^{n} \ell_{\text{cs}}(y^i, \langle \mathbf{W}, \mathbf{x}^i \rangle),$$

Let $\tilde{\xi}_i := \ell_{\text{cs}}(y^i, \left\langle \tilde{\mathbf{W}}, \mathbf{x}^i \right\rangle)$. It is easy to see that $\left\{\tilde{\mathbf{W}}, \{\tilde{\xi}_i\}\right\}$ satisfy all the constraints of (P_1). Since $\{\mathbf{W}, \{\xi_i\}\}$ satisfies the conditions of (P_1) (since it is an assumed optimum for (P_1)) we must have $\xi_i \geq \ell_{\text{cs}}(y^i, \left\langle \mathbf{W}, \mathbf{x}^i \right\rangle)$ (see lemma above). The above give us

$$\sum_{k=1}^{K} \left\|\tilde{\mathbf{w}}^k\right\|_2^2 + \sum_{i=1}^{n} \tilde{\xi}_i < \sum_{k=1}^{K} \left\|\mathbf{w}^k\right\|_2^2 + \sum_{i=1}^{n} \xi_i$$

which means that $\left\{\tilde{\mathbf{W}}, \{\tilde{\xi}_i\}\right\}$ is a better solution to (P_1) than $\{\mathbf{W}, \{\xi_i\}\}$. This again contradicts our earlier assumption that $\{\mathbf{W}, \{\xi_i\}\}$ is an optimum for (P_1). This means that $\mathbf{W}$ must be an optimum for (P_2).

Problem 1.5 (Sub-gradient Computation)**.** Consider the following function, where $(\mathbf{x}^i, y^i)_{i=1,\ldots,n}$, where $\mathbf{x}^i \in \mathbb{R}^d$ and $y^i \in \{-1, 1\}$ i.e. binary Rademacher labels.

$$f(\mathbf{w}) = \sum_{i=1}^{n} [1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle]_+$$

Suppose I construct a vector $\mathbf{g}$ as follows $\mathbf{g} = \sum_{i=1}^{n} \mathbf{h}^i$, where

$$\mathbf{h}^i = \begin{cases} -y^i \cdot \mathbf{x}^i & \text{if } \; y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle < 1 \\ \mathbf{0} & \text{if } \; y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \geq 1, \end{cases}$$

then show that $\mathbf{g} \in \partial f(\mathbf{w})$ i.e. $\mathbf{g}$ is a member of the subdifferential of f at $\mathbf{w}$. Recall that to show this, you have to show that for every $\mathbf{w}' \in \mathbb{R}^d$, $f(\mathbf{w}') \geq f(\mathbf{w}) + \langle \mathbf{g}, \mathbf{w}' - \mathbf{w} \rangle$. (5 marks)

Solution. For any $i \in [n]$, denote $f_i(\mathbf{w}) := [1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle]_+$. I will show that for every $i \in [n]$, we have, for every $\mathbf{w}' \in \mathbb{R}^d$, $f_i(\mathbf{w}') \geq f_i(\mathbf{w}) + \left\langle \mathbf{h}^i, \mathbf{w}' - \mathbf{w} \right\rangle$. Adding all these inequalities will then show for every $\mathbf{w}' \in \mathbb{R}^d$, $f(\mathbf{w}') \geq f(\mathbf{w}) + \langle \mathbf{g}, \mathbf{w}' - \mathbf{w} \rangle$ by linearity of the inner/dot product and the fact that $\mathbf{g} = \sum_{i=1}^{n} \mathbf{h}^i$. We will consider two cases.

In case 1, we consider i such that $y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \geq 1$. In such a situation, we know by definition that $f_i(\mathbf{w}) = 0$. However, by definition of the positive-part function $[\cdot]_+$, we know that $f_i(\mathbf{w}') := [1 - y^i \left\langle \mathbf{w}', \mathbf{x}^i \right\rangle]_+ \geq 0$. This gives us

$$f_i(\mathbf{w}') \geq 0 = f_i(\mathbf{w}) = f_i(\mathbf{w}) + \left\langle \mathbf{0}, \mathbf{w}' - \mathbf{w} \right\rangle$$

This finishes case 1.

In case 2, we consider i such that $y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle < 1$. In such a situation, we know by definition of the positive-part function that $f_i(\mathbf{w}) = 1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle$. In this case we also have $\mathbf{h}^i = -y^i \cdot \mathbf{x}^i$. Also, by definition of the positive-part function $[\cdot]_+$, we know that for any $v \in \mathbb{R}$, we have $[v]_+ \geq v$. This gives us

$$f_i(\mathbf{w}') = [1 - y^i \left\langle \mathbf{w}', \mathbf{x}^i \right\rangle]_+ \geq 1 - y^i \left\langle \mathbf{w}', \mathbf{x}^i \right\rangle = 1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle + \left\langle -y^i \cdot \mathbf{x}^i, \mathbf{w}' - \mathbf{w} \right\rangle = f_i(\mathbf{w}) + \left\langle \mathbf{h}^i, \mathbf{w}' - \mathbf{w} \right\rangle$$

This finishes case 2.