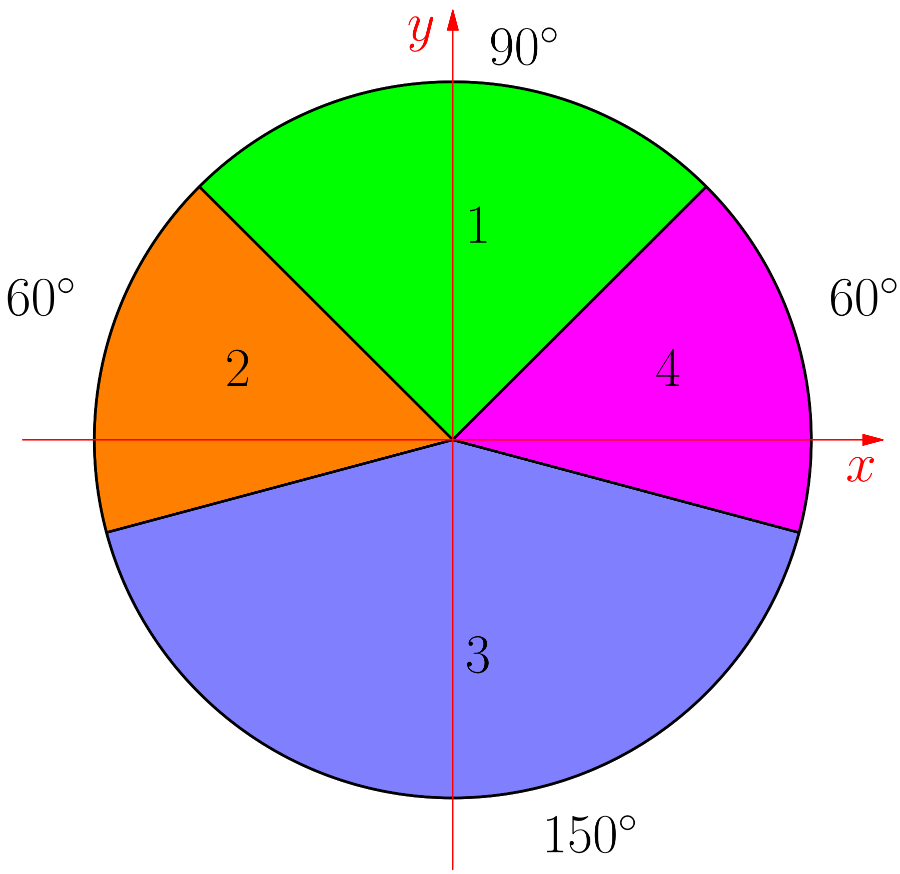
Machine Learning – Brett Bernstein

Week 7 Lecture: Concept Check Exercises

Multiclass

1. Let $\mathcal{X} = \mathbb{R}^2$ and $\mathcal{Y} = \{1, 2, 3, 4\}$, with X uniformly distributed on $\{x \mid \|x\|_2 \leq 1\}$. Given X, the value of Y is determined according to the following image, where green is 1, orange is 2, blue is 3, and magenta is 4.



For the problems below we are using the 0-1 loss.

(a) Consider the multiclass linear hypothesis space

$$\mathcal{F} = \{f \mid f(x) = \underset{i \in \{1,2,3,4\}}{\arg\max}\, w_i^T x\},$$

where each f is determined by $w_1, w_2, w_3, w_4 \in \mathbb{R}^2$. Give $f_{\mathcal{F}}$, a decision function minimizing the risk over $\mathcal{F}$, by specifying the corresponding w_1, w_2, w_3, w_4. Then give $R(f_{\mathcal{F}})$.

(b) Now consider the restricted hypothesis space

$$\mathcal{F}_1 = \{f \mid f(x) = \arg\max_{i\in\{1,2,3,4\}} w_i^T x, \|w_1\| = \|w_2\| = \|w_3\| = \|w_4\| = 1\}.$$

Consider the decision function $f \in \mathcal{F}_1$ with w_1, w_2, w_3, w_3 set to the angle bisectors of the corresponding regions. Give $R(f)$.

(c) Next consider the class-sensitive version of $\mathcal{F}$:

$$\mathcal{F}_2 = \{f \mid f(x) = \arg\max_{i\in\{1,2,3,4\}} w^T \Psi(x,i)\},$$

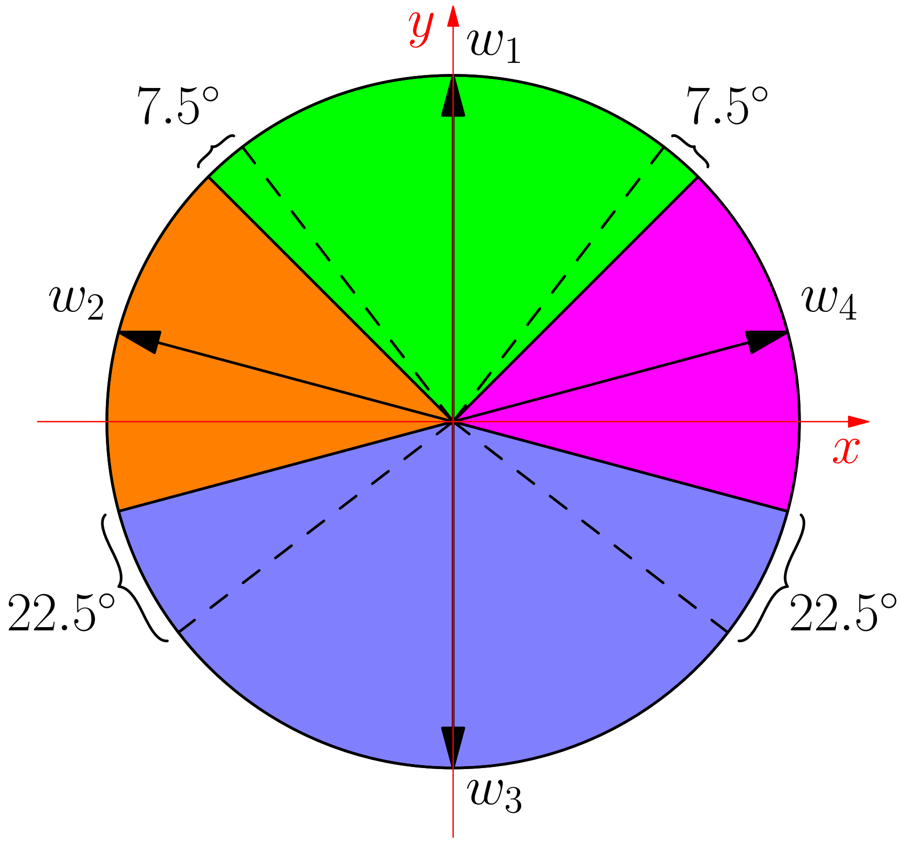
where $w \in \mathbb{R}^D$ and $\Psi : \mathbb{R}^2 \times \{1,2,3,4\} \to \mathbb{R}^D$. Give w, Ψ corresponding to $f_{\mathcal{F}_2}$, the decision function minimizing the risk over $\mathcal{F}_2$.

Solution.

(a) Let $w_1 = (0,1)^T$, $w_2 = (-1,0)^T$, $w_3 = (0,-c)^T$, $w_4 = (1,0)^T$, where $c = \cot\frac{\pi}{12} = 2+\sqrt{3}$. The corresponding risk is 0. To see how c was computed, consider the boundary between the magenta and blue regions. The division occurs along the vector $(\cos(\pi/12), -\sin(\pi/12))$. Note that

$$w_4^T(\cos(\pi/12), -\sin(\pi/12)) = \cos(\pi/12) = w_3^T(\cos(\pi/12), -\sin(\pi/12)).$$

(b) We have $w_1 = (0,1)$, $w_3 = (0,-1)$, $w_2 = (-\cos(\pi/2), \sin(\pi/12))$, $w_4 = (\cos(\pi/12), \sin(\pi/12))$. This gives the image below.



The dashed lines above are the boundaries of the 4 regions. The resulting risk is $(7.5 + 7.5 + 22.5 + 22.5)/360 = 1/6$.

2. Recall that the standard (featurized) SVM objective is given by

$$J_1(w) = \frac{1}{2}\|w\|_2^2 + \frac{C}{n}\sum_{i=1}^{n}[1 - y_i w^T \varphi(x_i)]_+.$$

The 2-class multiclass SVM objective is given by

$$J_2(w) = \frac{1}{2}\|w\|_2^2 + \frac{C}{n}\sum_{i=1}^{n}\max_{y \neq y_i}[1 - m_{i,y}(w)]_+,$$

where $m_{i,y}(w) = w^T\Psi(x_i, y_i) - w^T\Psi(x_i, y)$. Give a Ψ (in terms of φ) so that multiclass with 2 classes $\{-1, +1\}$ is equivalent to our standard SVM objective.

Solution. Let $\Psi(x, y) = \frac{1}{2}yx$ for $y \in \{-1, +1\}$. Then we have, for $y \neq y_i$,

$$1 - m_{i,y}(w) = 1 - (w^T x_i y_i - w^T x_i y)/2 = \begin{cases} 1 + w^T x_i & \text{if } y_i = -1, \\ 1 - w^T x_i & \text{if } y_i = +1. \end{cases}$$

This gives $1 - m_{i,y}(w) = 1 - y_i w^T \varphi(x_i)$.

3. Suppose you trained a decision function f from the hypothesis space $\mathcal{F}$ given by

$$\mathcal{F} = \{f \mid f(x) = \arg\max_{i \in \{1,\ldots,k\}} w^T \psi(x,i)\}.$$

Give pseudocode showing how you would use f to forecast the class of a new data point x.

Solution.

(a) Evaluate $w^T\psi(x,i)$ for $i = 1,\ldots,k$.

(b) Forecast the value i that gives the largest $w^T\psi(x,i)$ value.

4. Consider a multiclass SVM with objective

$$J(w) = \frac{1}{2}\|w\|_2^2 + \frac{C}{n}\sum_{i=1}^{n} \max_{y \neq y_i}[1 - m_{i,y}(w)]_+,$$

where $m_{i,y}(w) = w^T\Psi(x_i, y_i) - w^T\Psi(x_i, y)$. Assume $\mathcal{Y} = \{1,\ldots,k\}$, $\mathcal{X} = \mathbb{R}^d$, $w \in \mathbb{R}^D$ and $\psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^D$. Give a kernelized version of the objective.

Solution. Let $X \in \mathbb{R}^{nk \times D}$ matrix that has each $\Psi(x_i, y)^T$ as rows for each $i = 1,\ldots,n$ and $y = 1,\ldots,k$. More precisely, $\Psi(x_i, y)^T$ will be in row $(i-1)k + y$ of X. By the representer theorem, a solution, if it exists, must have the form $w^* = X^T\alpha$. Let $XX^T = K$, the Gram matrix. Then we have

$$m_{i,y}(w) = w^T\Psi(x_i, y_i) - w^T\Psi(x_i, y) = (K\alpha)_{(i-1)k+y_i} - (K\alpha)_{(i-1)k+y},$$

and $\|w\|_2^2 = \alpha^T K\alpha$. Substituting we have

$$J(\alpha) = \frac{1}{2}\alpha^T K\alpha + \frac{C}{n}\sum_{i=1}^{n} \max_{y \neq y_i}(1 - ((K\alpha)_{(i-1)k+y_i} - (K\alpha)_{(i-1)k+y}))_+.$$

Note that the Gram matrix K is $nk \times nk$, and thus can be infeasible to store or compute for nk large.