Title

David S. Rosenberg

1 Perceptron

The **perceptron loss** is given by

$$\ell(\hat{y}, y) = \max\{0, -\hat{y}y\}.$$

And consider the of linear functions $\mathcal{H} = \left\{f \mid f(x) = w^T x,\, w \in \mathbf{R}^d\right\}$.

1. [1] Suppose we have a linear function $f(x) = w^T x$, for some $w \in \mathbf{R}^d$. Geometrically, we say that the hyperplane $H = \{x \mid f(x) = 0\}$ separates the dataset $\mathcal{D} = ((x_1, y_1), \ldots, (x_n, y_n)) \in \mathbf{R}^d \times \{-1, 1\}$ if all x_i corresponding to $y_i = -1$ are strictly on one side of H, and all x_i corresponding to $y_i = 1$ are strictly on the other side of H. ("Strictly" here means that no x_i's lie on H.) Give a mathematical formulation of the necessary and sufficient conditions for $f(x) = w^T x$ to separate $\mathcal{D}$.
   SOLUTION:
   $$y_i f(x_i) > 0 \; \forall i \in \{1, \ldots, n\}$$

2. [1] In the homework we showed that if our prediction function $f(x) = w^T x$ separates a dataset $\mathcal{D}$, then the average perceptron loss on $\mathcal{D}$ is 0. The converse is not true: we many have average perceptron loss 0, but $f(x)$ may not properly separate $\mathcal{D}$. Explain why.
   SOLUTION: We have 0 loss even if x_i lies on h. An extreme example is $w = 0$, which has 0 loss.

2 Regularized Perceptron

Consider a hypothesis space of linear functions $\mathcal{H} = \left\{f \mid f(x) = w^T x,\, w \in \mathbf{R}^d\right\}$. Let $\ell(\hat{y}, y) = \max\{0, -\hat{y}y\}$ be the Perceptron loss. Consider choosing w minimizingsolving the following regularized empirical risk objective

$$J(w) = \frac{1}{2}\|w\|^2 + \frac{c}{n}\sum_{i=1}^{n} \max\left\{0, -y_i w^T x_i\right\}$$

for some $c > 0$.

1. Let $J_1(w; x, y) = \frac{1}{2}\|w\|^2 + c \max\left\{0, -y w^T x\right\}$. Give a subgradient g of $J_1(w; x, y)$ with respect to w. The subgradient will be a function of x, y, c, and w.
   Solution:

$$g = \begin{cases} -cyx + w & \text{for } yw^T x < 0 \\ w & \text{for } yw^T x \geq 0. \end{cases}$$

2. Give pseudocode or otherwise explain how you would use stochastic subgradient descent to find a minimizer w^* of $J(w)$. You need to specify your approach to the step size, but you do not have to specify a stopping criterion, though you may if you like.

3. $\min_{w \in \mathbf{R}^d} J(w)$ is an unconstrained minimization problem with a non-differentiable objective function. Rewrite it as a constrained optimization problem with a differentiable objective. [Hint: You may want to introduce new variables as we did for the SVM.]
   Solution:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|^2 + \frac{c}{n}\sum_{i=1}^{n} \xi_i \\ \text{such that} \quad & -\xi_i \leq 0 \ \forall i \\ & -\xi_i - y_i w^T x_i \leq 0 \ \forall i \end{aligned}$$

4. Write the Lagrangian for the optimization problem in (3) and the dual optimization problem.
   SOLUTION:

$$\begin{aligned} L(w, \xi, \lambda, \alpha) &= \frac{1}{2}\|w\|^2 + \frac{c}{n}\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \lambda_i \xi_i - \sum_{i=1}^{n} \alpha_i \left(\xi_i + y_i w^T x_i\right) \\ &= \frac{1}{2} w^T w + \sum_{i=1}^{n} \xi_i \left(\frac{c}{n} - \lambda_i - \alpha_i\right) - \sum_{i=1}^{n} \alpha_i y_i w^T x_i \end{aligned}$$

The dual optimization problem is

$$\sup_{\alpha,\lambda\succeq 0} \inf_{w,\xi} L(w,\xi,\lambda,\alpha).$$

First order conditions for the inner minimization we have

$$\begin{aligned}
\partial_w L(w,\xi,\lambda,\alpha) = 0 \quad &\Longleftrightarrow \quad w = \sum_{i=1}^n \alpha_i y_i x_i \\
\partial_{\xi_i} L = 0 \quad &\Longleftrightarrow \quad \frac{c}{n} - \lambda_i - \alpha_i = 0 \Longleftrightarrow \alpha_i + \lambda_i = \frac{c}{n}
\end{aligned}$$

Substituting these conditions back in to L we get

$$\begin{aligned}
L(w,\xi,\lambda,\alpha) &= \frac{1}{2}\sum_{i,j=1}^n \alpha_i\alpha_j y_i y_j x_i^T x_j - \sum_{i=1}^n\sum_{j=1}^n \alpha_i y_i \alpha_j y_j x_j^T x_i \\
&= -\frac{1}{2}\sum_{i,j=1}^n \alpha_i\alpha_j y_i y_j x_i^T x_j
\end{aligned}$$

Note that when we don't have $\frac{c}{n} - \lambda_i - \alpha_i = 0$, the $\inf_\xi$ is $-\infty$. Thus the dual function is

$$g(\alpha,\lambda) = \begin{cases} -\frac{1}{2}\sum_{i,j=1}^n \alpha_i\alpha_j y_i y_j x_i^T x_j & \text{for } \alpha_i + \lambda_i = \frac{c}{n} \text{ for all } i \\ -\infty & \text{otherwise} \end{cases}$$

So the dual optimization problem is given by

$$\begin{aligned}
\sup_{\alpha,\lambda} \quad & -\frac{1}{2}\sum_{i,j=1}^n \alpha_i\alpha_j y_i y_j x_i^T x_j \\
\text{s.t.} \quad & \alpha_i + \lambda_i = \frac{c}{n} \\
& \alpha_i \geq 0 \ \lambda_i \geq 0
\end{aligned}$$

We can eliminate λ and write this as

$$\begin{aligned}
\sup_{\alpha,\lambda} \quad & -\frac{1}{2}\sum_{i,j=1}^n \alpha_i\alpha_j y_i y_j x_i^T x_j \\
\text{s.t.} \quad & \alpha_i \in \left[0, \frac{c}{n}\right]
\end{aligned}$$

By Slater's condition, we have strong duality if we can find a feasible point. This is given by $w = 0$ and $\xi_i = 1$ for all i. Thus complementary slackness,

$$\begin{aligned} \lambda_i^* \xi_i^* = \left(\frac{c}{n} - \alpha_i^*\right) \xi_i^* &= 0 \\ \alpha_i^* \left(\xi_i^* + y_i f^*(x_i)\right) &= 0 \end{aligned}$$

Recall that ξ_i^* is the perceptron loss. So $\xi_i^* = 0 \iff y_i f^*(x_i) \geq 0$.

- $\alpha_i^* \in [0, c/n)$ implies $\xi_i^* = 0$ by the first condition. The second condition implies $\xi_i^* = -y_i f^*(x_i) = 0$. So $\alpha_i^* = 0$ implies the predictions are 0.
- If $y_i f^*(x_i) < 0$ then we have a loss, so $\xi_i^* > 0$. So by the first equation, $\alpha_i^* = \frac{c}{n}$.
- If $y_i f^*(x_i) > 0$ then we do not have a loss, so $\xi_i^* = 0$. Then second condition implies that $\alpha_i^* = 0$.
- So $\alpha_i^* = c/n$ implies $y_i f^*(x_i) \leq 0$.

We summarize these results below:

$$\begin{aligned} \alpha_i^* \in [0, \frac{c}{n}) &\implies f^*(x_i) = 0 \\ \alpha_i^* = \frac{c}{n} &\implies y_i f^*(x_i) \leq 0 \end{aligned}$$

$$\begin{aligned} y_i f^*(x_i) < 0 &\implies \alpha_i^* = \frac{c}{n} \\ y_i f^*(x_i) > 0 &\implies \alpha_i^* = 0 \end{aligned}$$

Well wait – this means that if everything is correctly separated, then $w = 0$. In fact, $w = 0$ is always the optimal solution. This problem just got really boring.