MACHINE LEARNING CLASS NOTES 9-25-12

PROF. DAVID SONTAG

1 Kernel methods & optimization

One example of a kernel that is frequently used in practice and which allows for highly non-linear discriminant functions is the Gaussian kernel,

$$\exp\left(\frac{-\|\vec{x}-\vec{y}\|^2}{2\sigma^2}\right)$$

For the Gaussian kernel, $k(\vec{x},\vec{x})=1$ for any vector $\vec{x}$, and $k(\vec{x},\vec{y})\approx 0$ if x is very different from y. Thus, a kernel function can be interpreted as a similarity function. However, not just any similarity function is a *valid* kernel. In particular, recall that (by definition) $k(\vec{x},\vec{y})$ is a **valid kernel** if and only if $\exists\phi:\mathcal{X}\rightarrow\mathbb{R}^d$ s.t. $k(\vec{x},\vec{y})=\phi(\vec{x})\cdot\phi(\vec{y})$. One consequence of this is that kernel functions must be symmetric, since $\phi(\vec{x})\cdot\phi(\vec{y})=\phi(\vec{y})\cdot\phi(\vec{x})$.

Today's lecture will explore these requirements of kernel functions in more depth, culmunating with Mercer's theorem. Together, these requirements provide a mathematical foundation for kernel methods, ensuring both that there is a sensible feature vector representation for every data point and that the support vector machine (SVM) objective has a unique global optimum and is easy to optimize.

1.1 Background from linear algebra

A matrix $M\in\mathbb{R}^d\times\mathbb{R}^d$ is said to be **positive semi-definite** if $\forall z\in\mathbb{R}^d$, $z^TMz\geq 0$. For example, suppose $M=I$. Then,

$$z^TIz=\sum_{i=1}^{d}\sum_{j=1}^{d}z_iz_jI_{ij}=\sum_{i=1}^{d}z^2,$$

which is always ≥ 0. Thus, the identify matrix is positive semi-definite. Next we review several concepts from linear algebra, and then use these to give an alternative definition of positive semi-definite (PSD) matrices.

Suppose we find a vector $\vec{v}$ and a value λ such that $M\vec{v}=\lambda\vec{v}$. We call $\vec{v}$ an **eigenvector** of the matrix M, and λ an **eigenvalue**. A matrix M can be shown to be PSD if and only if M has all non-negative eigenvalues. We will now show one of the directions ($\Leftarrow$). To see this, first write $M=V\Lambda V^T$, where Λ is a matrix with the eigenvalues along the diagonal (zero off diagonal) and V is the matrix of eigenvectors:

$$M=V\begin{bmatrix}\lambda_1 & ... & 0 & 0\\ 0 & \lambda_2 & 0 & 0\\ 0 & 0 & ... & 0\\ 0 & 0 & 0 & \lambda_d\end{bmatrix}V^T$$

Next, we split Λ in two,

$$M = \left(V \begin{bmatrix} \sqrt{\lambda_1} & ... & 0 & 0 \\ 0 & \sqrt{\lambda_2} & 0 & 0 \\ 0 & 0 & ... & 0 \\ 0 & 0 & 0 & \sqrt{\lambda_d} \end{bmatrix} \right) \left(\begin{bmatrix} \sqrt{\lambda_1} & ... & 0 & 0 \\ 0 & \sqrt{\lambda_2} & 0 & 0 \\ 0 & 0 & ... & 0 \\ 0 & 0 & 0 & \sqrt{\lambda_d} \end{bmatrix} V^T \right) = UU^T.$$

Letting $v = z^T U$, since $vv^T = v \cdot v \geq 0$ we have that $(z^T U)(U^T z) = z^T M z \geq 0$, showing that M is positive semi-definite (we used the fact that the eigenvalues were non-negative when taking their square root).

1.2 Mercer's Theorem

For a training set $S = \{\vec{x_i}\}$ and a function $k(\vec{u}, \vec{v})$, the **kernel matrix** (also called the Gram matrix) K_S is the matrix of dimension $|S| \times |S|$ where $(K_S)_{ij} = k(\vec{x}_i, \vec{x}_j)$.

Theorem 1 (Mercer's theorem)**.** *$k(\vec{u}, \vec{v})$ is a valid kernel if and only if the corresponding kernel matrix is PSD for all training sets $S = \{\vec{x_i}\}$.*

Proof. ($\Rightarrow$) Since $k(\vec{u}, \vec{v})$ is a valid kernel, it has a corresponding feature map ϕ such that $k(\vec{u}, \vec{v}) = \phi(\vec{u}) \cdot \phi(\vec{v})$. Thus, the kernel matrix K_s has entries $(K_S)_{ij} = \phi(\vec{x}_i) \cdot \phi(\vec{x}_j)$. Let V be the matrix $\begin{bmatrix} \phi(x_1) & ... & \phi(x_n) \end{bmatrix}$, where we treat $\phi(x_i)$ as a column vector. Then, we have $K_S = V^T V$. However, thsi shows that K_S must be positive semi-definite, because for any vector $z \in \mathbb{R}^{|S|}$, $(z^T V^T)(Vz) \geq 0$.

($\Leftarrow$) Let S be the set of all possible data points (we will assume that it is finite). Since the corresponding kernel matrix K_S is positive semi-definite, it has non-negative eigenvalues and can be factored as $K_S = UU^T$. Let $\phi(x_i) = u_i$, where u_i is the i'th row of U. This gives the feature mapping for x_i such that $k(x_i, x_j) = u_i \cdot u_j$. □

Mercer's theorem guarantees for us that the kernel matrix is positive semi-definite. As we show in the next section, this will guarantee that the SVM dual objective is concave, which means that it is easy to optimize.

1.3 Convexity

A **set** $X \subseteq \mathbb{R}^d$ is a **convex** set if for any $\vec{x}, \vec{y} \in X$ and $0 \leq \alpha \leq 1$,

$$\alpha\vec{x} + (1 - \alpha)\vec{y} \in X$$

Informally, if for any two points $\vec{x}$, $\vec{y}$ that are in the set every point on the line connecting $\vec{x}$ and $\vec{y}$ is also included in the set, then the set is convex. See Figure 1 for examples of non-convex and convex sets.

A **function** $f : \mathcal{X} \to \mathbb{R}$ is **convex** for a convex set $\mathcal{X}$ if $\forall \vec{x}, \vec{y} \in \mathcal{X}$ and $0 \leq \alpha \leq 1$,

$$f(\alpha\vec{x} + (1 - \alpha)\vec{y}) \leq \alpha f(\vec{x}) + (1 - \alpha) f(\vec{y}) \tag{1}$$

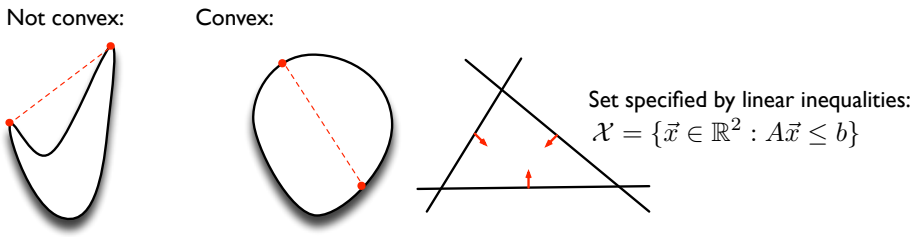

Figure 1: Illustration of a non-convex and two convex sets in $\mathbb{R}^2$.

Informally, a function is convex if the line between any two points on the curve always upper bounds the function. We call a function **strictly convex** if the inequality in Eq. 1 is a strict inequality. See See Figure 2 for examples of non-convex and convex functions. A function $f(x)$ is **concave** is $-f(x)$ is convex. Importantly, it can be shown that strictly convex functions always have a unique minima.

For a function $f(x)$ defined over the real line, one can show that $f(x)$ is convex if and only if $\frac{d^2}{dx^2}f \geq 0 \; \forall x$. Just as before, strict convexity occurs when the inequality is strict. For example, consider $f(x) = x^2$. The first derivative of $f(x)$ is given by $\frac{d}{dx}f = 2x$ and its second derivative by $\frac{d^2}{dx^2}f = 2$. Since this is always strictly greater than 0, we have proven that $f(x) = x^2$ is strictly convex. As a second example, consider $f(x) = \log(x)$. The first derivative is $\frac{d}{dx}f = \frac{1}{x}$, and its second derivative is given by $\frac{d^2}{dx^2}f = -\frac{1}{x^2}$. Since this is negative for all $x > 0$, we have proven that $\log(x)$ is a *concave* function over $\mathbb{R}_+$.

This matters because optimization for convex functions is easy. In particular, one can show that nearly any reasonable optimization method, such as gradient descent (where one starts at arbitrary point, moves a little bit in the direction opposite to the gradient, and then repeats), is **guaranteed** to reach a global optimum of the function. Note that whereas the minimization of convex functions is easy, likewise, the maximization of concave functions is easy.

Finally, to generalize this second definition of convex functions to higher dimensions (i.e., $\mathcal{X} = \mathbb{R}^d$), we introduce the notion of the **Hessian matrix** of

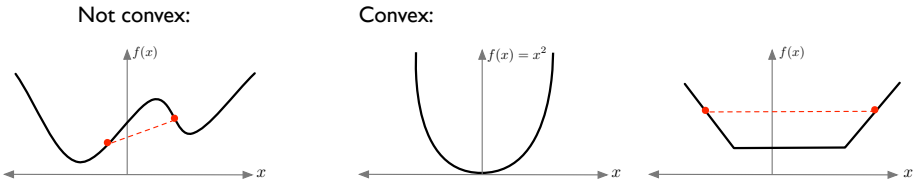

Figure 2: Illustration of a non-convex and two convex functions over $\mathcal{X} = \mathbb{R}$.

a function f,

$$\nabla^2 f(\vec{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_d^2} \end{bmatrix}$$

which is the matrix of dimension $d \times d$ with entries $(\nabla^2 f)_{ij}$ equal to the partial derivative of the function with respect to x_i and then with respect to x_j, denoted $\frac{\partial^2 f}{\partial x_i \partial x_j}$. Note that since the order of the partial derivatives does not matter, i.e. $\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$, the Hessian matrix is symmetric.

We are finally ready for our second definition of convex functions in higher dimension. A function $f : \mathcal{X} \to \mathbb{R}$ is **convex** for a convex set $\mathcal{X} \subseteq \mathbb{R}^d$ if and only if its Hessian matrix $\nabla^2 f(\vec{x})$ is positive semi-definite for all $\vec{x} \in \mathcal{X}$.

1.4 The dual SVM objective is concave

Recall the dual of the support vector machine (SVM) objective,

$$f(\vec{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j k(x_i, x_j) \tag{2}$$

The first partial derivative is given by

$$\frac{\partial f}{\partial \alpha_s} = 1 - \sum_{i \neq s} \alpha_i y_i y_s k(x_i, x_s) - \alpha_s k(x_s, x_s)$$

The second partial derivative is given by

$$\frac{\partial^2 f}{\partial \alpha_t \partial \alpha_s} = -y_t y_s k(x_t, x_s)$$

Let $\vec{y} \in \{-1, 1\}^n$ be the vector of assignments to the n data points (a column vector). We can then write the Hessian matrix $\nabla^2 f$ as $-\vec{y}^T K_S \vec{y}$, where K_S is the kernel matrix for the n data points. Since $k(\vec{u}, \vec{v})$ is a valid kernel, Mercer's theorem guarantees for us that K_S is positive semi-definite. As a result, we have that $\vec{z}^T K_S \vec{z} \geq 0$ for all vectors $\vec{z} \in \mathbb{R}^n$. We conclude that $-\vec{y}^T K \vec{y} \leq 0$, finishing our proof that the dual SVM objective is concave.

There are many approaches for minimizing $f(\vec{\alpha})$. One of the simplest such methods is called the sequential minimal optimization (SMO) algorithm, and is based on the concept of **block coordinate descent**. Coordinate descent is illustrated in Fig. 3 for a function defined on $\mathbb{R}^2$. An arbitrary starting point is chosen. Then, in each step, one coordinate (or, in general, a set of coordinates, called a block) is chosen and the function is minimized as much as possible with respect to that coordinate (keeping all other variables fixed to their current values).

The larger the blocks, the faster the convergence to the optimum solution. The blocks are typically chosen to be as large as possible such that minimizing

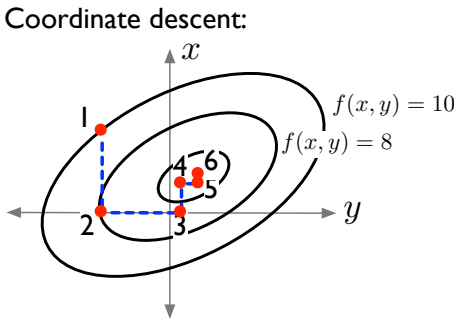

Figure 3: Illustration of coordinate descent on the function $f(x, y)$. Shown here are the level sets of the function. The numbers indicate the first point, the second point, etc., until the optimal solution is found.

the function with respect to these coordinates can be performed in closed form. For the dual SVM, because of the constraint $\sum_i y_i \alpha_i = 0$, the smallest block size that can be chosen is 2. The algorithm proceeds by choosing in each iteration α_i and α_j, then minimizing the function as much as possible with respect to these two variables.