Support Vector Machine (SVM)

- Binary Classification with linear models
- $y \in \{-1, 1\}$ and $x \in \mathbb{R}^d$

Constraint Optimization

Lagrange function

- $L(x, \lambda) := f(x) + \lambda h(x)$
  - λ is *Lagrange Multiplier*

Equality Constraint

- $\min\limits_{x} f(x) \quad s.t.\ \ h(x) = 0$
  - $\forall$ point x on the constraint surface $h(x) = 0$, we have $\nabla h(x)$ orthogonal to the surface
    - Because if $\nabla h(x)$ has tangent component, we can move along the direction to make $h(x) \neq 0$
  - For a local minimum x^*, $\nabla f(x^*)$ must be orthogonal to the surface
    - Generally, x^* is a local minimum $\Rightarrow \exists \lambda \quad s.t.\ \nabla f(x^*) + \lambda \nabla h(x^*) = 0$
    - (There may be some corner cases not satisfying this equation)
  - 可以说，极值点往往是限制平面与函数相切的点，那么此时该点处平面的梯度与函数的梯度就应该是相反的
- When we have K constraints:
  - $\min\limits_{x} f(x) \quad s.t.\ \ h_i(x) = 0,\ i \in \{1, \ldots, K\}$
  - then $L(x, \lambda) = f(x) + \sum_{i=1}^{K} \lambda_i h_i(x)$
- now we can say that:

$$x^* \ is\ a\ local\ minimum \Rightarrow \exists \lambda \begin{cases} \nabla_x L(x^*, \lambda) = 0 & (1) \\ \nabla_\lambda L(x^*, \lambda) = 0 & (2) \end{cases}$$

  - (1)说明 $\lambda f(x^*) + \sum_{i=1}^{n} \lambda_i \nabla h_i(x^*) = 0$
  - (2)说明 $h_i(x^*) = 0, \quad \forall i$

Inequality Constraint

- $\min\limits_{x} f(x) \quad s.t.\ \ g(x) \leq 0$
  - $\forall$ point x on the surface $g(x) = 0$, $\nabla g(x)$ must be orthogonal to the surface and *point out of* the region of $g(x) \leq 0$
  - For a local minimum x^*
    - the constraint is **active**:
      - if x^* on the surface $g(x) = 0$, then $-\nabla f(x^*)$ must have the same direction as $\nabla g(x^*)$
      - $\exists \mu > 0,\ \ s.t.\ \nabla f(x^*) + \mu \nabla g(x^*) = 0$
    - the constraint is **inactive**:
      - if x^* within the surface, we only need $\nabla f(x^*) = 0$
      - $\exists \mu = 0,\ \ s.t.\ \nabla f(x^*) + \mu \nabla g(x^*) = 0$
    - 显然地，如果函数的最小值点在平面内部，自然这个平面的约束就没有用了，极值点就是这个最小值点；而如果函数的最小值点在平面外，那么就和equality constraint一样，极值点在平面的边界上
    - 总结:

$$\exists \mu \geq 0,\ \ s.t.\ \nabla f(x^*) + \mu \nabla g(x^*) = 0 \begin{cases} \mu = 0 & inactive \\ \mu > 0 & active \end{cases}$$

K.K.T conditions

- $\min\limits_{x} f(x) \quad s.t.\ \ h_i(x) = 0,\ i \in \{1, \ldots, K\},\ \ g_j(x) \leq 0,\ j \in \{1, \ldots, L\}$
- $L(x, \lambda, \mu) = f(x) + \sum_{i=1}^{K} \lambda_i h_i(x) + \sum_{j=1}^{L} \mu_j g_j(x)$
- **K.K.T conditions**:

$$x^*\ is\ a\ local\ minimum \Rightarrow \begin{cases} \nabla_x L(x^*,\lambda,\mu)=0 & \\ h_i(x^*)=0 & \forall i \\ g_j(x^*)\le 0 & \forall j \\ \mu_j \ge 0 & \forall j \\ \mu_j g_j(x^*)=0 & \forall j \end{cases}$$

- 对于$\mu_j g_j(x^*)=0$，若x^*在边界处，显然有$g_j(x^*)=0$；若x^*在内部，则有$\mu=0$，所以该式恒等于0

Primal Form

- **Maximize Margin Criterion**
  - minimum distance to the hyperplane $w^Tx+b=0$, over all training data
  - *Structural Risk Minimization*
- x到$w^Tx+b=0$的距离为$\Delta x=\frac{w^Tx+b}{||w||}$
- we can define $\gamma_i := y_i\Delta x_i = \frac{y_i(w^Tx+b)}{||w||} \ge 0$ (as $y_i \in \{-1,1\}$)
- $\gamma = \underset{i=1,...,n}{min}\ \gamma_i = \underset{i=1,...,n}{min}\ \frac{y_i(w^Tx+b)}{||w||}$
  - this is called **geometric margin**
- now we need to maximize this distance:

$$\underset{w,b,\gamma}{max}\,\gamma \quad s.t. \frac{y_i(w^Tx_i+b)}{||w||} \ge \gamma, \quad \forall i$$

  - 但显然三个参数对于我们来说还是太多了，所以我们可以先假设找到了离超平面$w^Tx+b=0$最近的点x_0，此时有$\gamma_0=\frac{y_0(w^Tx_0+b)}{||w||}=\gamma$
- then we can update our object:

$$\underset{w,b}{max}\,\frac{y_0(w^Tx_0+b)}{||w||} \quad s.t.\ \ y_i(w^Tx_i+b)\ge y_0(w^Tx_0+b),\ \ \forall i$$

  - 我们可以发现，$w^Tx+b=0$是一个超平面，而$kw^Tx+kb=0,\ \forall k$都是同一个超平面，那么这样的话$y_0(w^Tx_0+b)$就可以是任意值(which is called **functional margin**)，所以我们可以干脆令$y_0(w^Tx_0+b)=1$，背后的逻辑是不管真实的γ是多少，我都可以找到一个k使上式等于1(同理，不管上式等于多少，我们也总能找到相应的w使得几何距离不变)，而且显然这个等式也是不影响γ的
- then we can say that:

$$\underset{w,b}{max}\,\frac{1}{||w||} \quad s.t.\ \ y_i(w^Tx_i+b)\ge 1, \quad \forall i$$

- now we have the **primal form** of SVM:

$$\underset{w,b}{min}\,\frac{1}{2}w^Tw \quad s.t.\ \ y_i(w^Tx_i+b)\ge 1, \quad \forall i$$

- 我们管x_0这样的用来确定超平面的向量叫做**支持向量**(**support vector**)，由此我们即可以看出支持向量机的命名缘由：这些向量支撑了超平面的选择，我们只需考虑这些支持向量，而不用管别的向量，就可以确定最终的超平面
- Convex Quadratic Programming
  - 这个函数是凸的，所以总能找到一个最优解，调包即可
  - Standard package: $O(d^3)$

Dual Form

- 2 problems for SVM:
  - may be not linearly separable
  - margin γ may be too small (there may be some outlier samples)
- 我们可以改写primal form，把条件写进函数中：

$$L(w,b,\alpha)=\frac{1}{2}||w||^2+\sum_{i=1}^{n}\alpha_i(1-y_i(w^Tx_i+b)) \quad (\alpha_i\ge 0)$$
$$p^*=\underset{w,b}{min}\,\underset{\alpha\ge 0}{max}\,L(w,b,\alpha) \quad \text{(primal form of SVM)}$$

  - if $1-y_i(w^Tx_i+b)>0$, then $\alpha\to\infty$ and $L\to\infty$
    - if p^* exists, $1-y_i(w^Tx_i+b)\le 0$
  - if $1-y_i(w^Tx_i+b)<0$, then $\alpha=0$
  - if $1-y_i(w^Tx_i+b)=0$, then $\alpha_i(1-y_i(w^Tx_i+b))=0$
  - this is part of **K.K.T conditions**!

- **Dual form of SVM**

$$d^* = \underset{\alpha \geq 0}{max}\ \underset{w,b}{min}\ L(w,b,\alpha)$$

Weak Duality

- $d^* \leq p^*$

$$d^* = \underset{\alpha \geq 0}{max}\ \underset{w,b}{min}\ L(w,b,\alpha) \leq \underset{\alpha \geq 0}{max}\ L(w,b,\alpha) \quad \forall w,b$$
$$d^* \leq \underset{w,b}{min}\ \underset{\alpha \geq 0}{max}\ L(w,b,\alpha) = p^*$$

- $p^* - d^*$ is called **duality gap**

Strong Duality

- $d^* = p^*$
- without duality gap!
- Slator's Condition
  - when objective is **convex** and constraints are **linear**, then $d^* = p^*$
  - our function just fits these conditions!
- So now we can solve the dual problem to solve the primal problem
  - if we have the solution of primal problem w_p, b_p and the solution of dual problem α_d

$$p^* = \underset{w,b}{min}\ \underset{\alpha \geq 0}{max}\ L(w,b,\alpha) = \underset{\alpha \geq 0}{max}\ L(w_p,b_p,\alpha)$$
$$\geq L(w_p,b_p,\alpha_d) \geq \underset{w,b}{min}\ L(w,b,\alpha_d) = \underset{\alpha \geq 0}{max}\ \underset{w,b}{min}\ L(w,b,\alpha) = d^*$$

  - 根据Slator's Condition，这里我们就可以直接取等了

Dual problem

- 我们首先来看函数的内层$\underset{w,b}{min}\ L(w,b,\alpha)$
  - 由于L是凸函数，所以我们可以通过找导函数的零点来求最小值

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0$$
$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0$$

  - 所以我们有：

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i = 0$$
$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

  - 可以发现这其实也是K.K.T conditions中的$\nabla_{w,b} L(w,b,\alpha) = 0$
- Substitute w into L:

$$L(w,b,\alpha) = \frac{1}{2}(\sum_{i=1}^{n} \alpha_i y_i x_i)^T(\sum_{j=1}^{n} \alpha_j y_j x_j) + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i y_i x_i^T(\sum_{j=1}^{n} \alpha_j y_j x_j)$$
$$= \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

- now we need to solve the **dual problem of SVM**:

$$max \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j$$
$$s.t. \quad \alpha_i \geq 0 \quad \forall i, \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

- **SMO**(Sequential Minimal Optimization)

- every time select α_i, α_j, fix remaining $n-2$ variables
  - $\alpha_i y_i + \alpha_j y_j = -\sum_{k\neq i,j} \alpha_k y_k$ represent α_j with α_i and solve problem w.r.t α_i
  - iterate until convergence
- When we have solution α^* for d^*, how to get w^*, b^*:

$$w^* = \sum_{i=1}^{n} \alpha_i y_i x_i$$
$$b^* = \frac{1}{y_i} - {w^*}^T x_i$$
$$(x_i, y_i) \text{ is a support vector}$$

- How to find a support vector:
  - recall: $\begin{cases} 1 - y_i(w^T x_i + b) = 0 & \Rightarrow & \alpha_i > 0 & \text{active} \\ 1 - y_i(w^T x_i + b) > 0 & \Rightarrow & \alpha_i = 0 & \text{inactive} \end{cases}$
  - then: $\begin{cases} \text{if } \alpha_i^* = 0 & \Rightarrow & (x_i, y_i) \text{is not a support vector} \\ \text{if } \alpha_i^* > 0 & \Rightarrow & (x_i, y_i) \text{is a support vector} \end{cases}$
  - we can save **only** the α_i^* of support vectors to save the model
  - α相当于权重，只有被认为是支持向量的权重才会大于0，其他都置0，表示其他的向量对于超平面完全是没用的，可以不用考虑
- when we have a new test point x:

$$f(x) = \sum_{i\in \text{support vectors}} \alpha_i^* y_i x_i^T x_i + b^*$$

  - 对于一个新来的点，我们只要让其和支持向量做内积，看$f(x)$大于0还是小于0即可将其分类

Kernel Trick

- Define similarity between 2 points X, Z
  - **Linear Kernel** $K(X,Z) = X^T Z$
  - **Polynomial Kernel** $K(X,Z) = (X^T Z + 1)^P$
  - **Gaussian / RBF(Radial Basis Function) Kernel** $K(X,Z) = exp(-\frac{||X-Z||^2}{2\sigma^2})$
- e.g. $\Phi: \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$
  - $\Phi(X)$ maps X to higher dimension
  - $X = (X_1, X_2)^T \in \mathbb{R}^2 \rightarrow \Phi(X) = (1, X_1, X_2, X_1^2, X_2^2, X_1X_2) \in \mathbb{R}^6$
- 很多时候，当数据线性不可分时，我们可以通过给数据升维，使得数据变得线性可分
- **Kernel Trick**:
  - a way to get $\Phi(X)^T \Phi(Z)$ directly in original space, without explicitly mapping data to high-dimension first
  - 我们虽然希望可以用高维数据进行计算，但实际上我们并不需要直接去求他们的内积，而是直接使用上述的核函数即可，因为这些核函数就可以写成一些高维数据的内积
  - **这样我们将上述dual problem中的$x_i^T x_j$部分替换为$K(x_i, x_j)$即可**
- e.g. $X = (X_1, X_2), Z = (Z_1, Z_2)$
  $(X^T Z + 1)^2 = (X_1 X_2 + Z_1 Z_2 + 1)^2 = (1, \sqrt{2}X_1, \sqrt{2}X_2, X_1^2, X_2^2, \sqrt{2}X_1X_2)^T (1, \sqrt{2}X_1, \sqrt{2}X_2, X_1^2, X_2^2, \sqrt{2}X_1X_2)$
- Gaussian Kernel

$$K(X,Z) = exp(-\frac{||X||^2}{2\sigma^2}) exp(-\frac{||Z||^2}{2\sigma^2}) exp(-\frac{X^T Z}{\sigma^2})$$

  and we know that:

$$exp(-\frac{X^T Z}{\sigma^2}) = 1 + \frac{X^T Z}{\sigma^2} + \frac{1}{2!}(\frac{X^T Z}{\sigma^2})^2 + \frac{1}{3!}(\frac{X^T Z}{\sigma^2})^3 + \dots$$
$$= \sum_{p=0}^{\infty} \frac{1}{p!}(\frac{X^T Z}{\sigma^2})^p$$

  - union of polynomial kernels from $p = 0$ to ∞
  - implicitly map data to infinite-dimension space and do inner product
  - 这里的σ可以自行控制，当σ大时，泰勒多项式会很快趋近于0，即此时映射成的向量维数不会很高；而当σ小时，则能够泰勒展开较多的项，此时就会映射成一个较高维的向量. 可以说，σ控制了映射维数的大小
  - $\sigma \rightarrow 0$时，$K(X,Z) \rightarrow 0$，此时类似$1-NN$，只有与X最近的点才会起作用
- d^*与p^*:
  - p^*计算w, b共$d+1$维，其对偶形式d^*则计算α共n维，一般来说n是很大的，但我们依然选择去用d^*计算，主要是考虑到了高斯核函数，根据上述的泰勒展开，我们可以发现理论上这个核函数可以把原始向量映射到一个**无穷维的空间**中，那么显然此时w, b也可以无限大，再用p^*来计算就不那么划算了

Soft-margin SVM

- Slack variables to handle outliers
  - previously, we require $y_i(w^Tx_i+b)\geq 1$. Now, we introduce slack variables $\xi_i \geq 0$ and require $y_i(w^Tx_i+b)\geq 1-\xi_i$
  - must restrict ξ_i to be small
  - ξ_i 表示样本最多可以与$w^Tx_i+b=0$**偏离的程度**

Unconstrained Form

- Soft-margin SVM

$$
\begin{aligned}
& \underset{w,b,\xi}{min}\frac{1}{2}||w||^2+c\sum_{i=1}^{n}\xi_i \\
s.t. \quad & y_i(w^Tx_i+b)\geq 1-\xi_i \\
& \xi_i \geq 0 \\
\text{化简后可得}\ & \underset{w,b}{min}\frac{1}{2}||w||^2+c\sum_{i=1}^{n}\max\{0,1-y_i(w^Tx_i+b)\}
\end{aligned}
$$

  - becomes an **unconstrained** optimization
  - $\max\{0,1-y_i(w^Tx_i+b)\}$ is called **Hinge Loss**
  - $c\rightarrow\infty\Rightarrow\xi_i=0\Rightarrow\text{hard margin}$

Dual Form

- Dual Form of soft-margin SVM

$$
\begin{aligned}
& \underset{\alpha,\beta\geq 0}{max}\,\underset{w,b,\xi}{min}L(w,b,\xi,\alpha,\beta) \\
& =\frac{1}{2}||w||^2+c\sum_{i=1}^{n}\xi_i+\sum_{i=1}^{n}\alpha_i(1-\xi_i-y_i(w^Tx_i+b))-\sum_{i=1}^{n}\beta_i\xi_i
\end{aligned}
$$

- Partial derivatives

$$
\begin{aligned}
& \frac{\partial L}{\partial w}=w-\sum_{i=1}^{n}\alpha_iy_ix_i=0 \\
& \frac{\partial L}{\partial b}=-\sum_{i=1}^{n}\alpha_iy_i=0 \\
& \frac{\partial L}{\partial \xi_i}=c-\alpha_i-\beta_i=0 \\
& \text{整理可得} \\
& max\,\frac{1}{2}(\sum_{i=1}^{n}\alpha_iy_ix_i)^T(\sum_{j=1}^{n}\alpha_jy_jx_j)+\sum_{i=1}^{n}\alpha_i-\sum_{i=1}^{n}\alpha_iy_ix_i^T(\sum_{j=1}^{n}\alpha_jy_jx_j) \\
= & max\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_jy_iy_jx_i^Tx_j \\
& s.t. \quad 0\leq\alpha_i\leq c, \quad \sum_{i=1}^{n}\alpha_iy_i=0
\end{aligned}
$$

  - $c\rightarrow\infty\Rightarrow\text{hard margin}$
  - c is small $\Rightarrow$ no individual point dominates the prediction(hyperplane) and is robust to outlier
  - c controls the **strength of regularization**