The Strategy So Far...

- Choose hypothesis space
- Construct loss function (ideally convex)
- Minimize loss to “learn” correct parameters
General Optimization

A mathematical detour, we’ll come back to SVMs soon!

$$\min_{x \in \mathbb{R}^n} f_0(x)$$

subject to:

$$f_i(x) \leq 0, \quad i = 1, \ldots, m$$
$$h_i(x) = 0, \quad i = 1, \ldots, p$$
General Optimization

\[
\min_{x \in \mathbb{R}^n} f_0(x)
\]

subject to:

\[
\begin{align*}
  f_i(x) &\leq 0, & i &= 1, \ldots, m \\
  h_i(x) &= 0, & i &= 1, \ldots, p
\end{align*}
\]

\(f_0\) is not necessarily convex
General Optimization

\[
\min_{x \in \mathbb{R}^n} f_0(x)
\]

subject to:

\[
\begin{align*}
  f_i(x) &\leq 0, & i = 1, \ldots, m \\
  h_i(x) &= 0, & i = 1, \ldots, p
\end{align*}
\]

Constraints do not need to be linear
Example

\[
\min_{x \in \mathbb{R}^3} x_1 \log x_1 + x_2 \log x_2
\]

subject to:

\[
\begin{align*}
    x_1 + x_2 &= 1 \\
    x_1 &\geq 0 \\
    x_2 &\geq 0
\end{align*}
\]
Example

\[
\min_{x \in \mathbb{R}^3} x_1 \log x_1 + x_2 \log x_2
\]

subject to:

\[
1 - x_1 - x_2 = 0
\]
\[
-x_1 \leq 0
\]
\[
-x_2 \leq 0
\]


Lagrangian

\[ L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \]

- Incorporate constraints into a new objective function
- \( \lambda \geq 0 \) and \( \nu \) are vectors of \textit{Lagrange multipliers}
- The Lagrange multipliers can be thought of as enforcing soft constraints
Example

\[ \min_{x \in \mathbb{R}^3} x_1 \log x_1 + x_2 \log x_2 \]

subject to:

\[ 1 - x_1 - x_2 = 0 \]
\[ -x_1 \leq 0 \]
\[ -x_2 \leq 0 \]

\[ L(x_1, x_2, \nu_1, \lambda_1, \lambda_2) = x_1 \log x_1 + x_2 \log x_2 + \nu_1 \cdot (1 - x_1 - x_2) - \lambda_1 x_1 - \lambda_2 x_2 \]
Duality

- Construct a dual function by minimizing the Lagrangian over the primal variables

\[ g(\lambda, \nu) = \inf_x L(x, \lambda, \nu) \]

- \( g(\lambda, \nu) = -\infty \) whenever the Lagrangian is not bounded from below for a fixed \( \lambda \) and \( \nu \)
Example

\[
\begin{align*}
\min_{x \in \mathbb{R}^3} & \quad x_1 \log x_1 + x_2 \log x_2 \\
\text{subject to:} & \quad 1 - x_1 - x_2 = 0 \\
& \quad -x_1 \leq 0 \\
& \quad -x_2 \leq 0
\end{align*}
\]

\[
L(x_1, x_2, \nu_1, \lambda_1, \lambda_2) = x_1 \log x_1 + x_2 \log x_2 + \nu_1 \cdot (1 - x_1 - x_2) - \lambda_1 x_1 - \lambda_2 x_2
\]

\[
\begin{align*}
\frac{\partial L}{\partial x_1} &= \log x_1 + 1 - \nu_1 - \lambda_1 = 0 \\
\Rightarrow x_1 &= \exp(\nu_1 + \lambda_1 - 1) \\
\frac{\partial L}{\partial x_2} &= \log x_2 + 1 - \nu_1 - \lambda_2 = 0 \\
\Rightarrow x_2 &= \exp(\nu_1 + \lambda_2 - 1)
\end{align*}
\]
Example

\[
\min_{x \in \mathbb{R}^3} x_1 \log x_1 + x_2 \log x_2
\]

subject to:

\[
1 - x_1 - x_2 = 0
\]
\[
-x_1 \leq 0
\]
\[
-x_2 \leq 0
\]

\[
L(x_1, x_2, \nu_1, \lambda_1, \lambda_2) = x_1 \log x_1 + x_2 \log x_2 + \nu_1 \cdot (1 - x_1 - x_2) - \lambda_1 x_1 - \lambda_2 x_2
\]

\[
g(\nu_1, \lambda_1, \lambda_2) = \exp(\nu_1 + \lambda_1 - 1) (\nu_1 + \lambda_1 - 1)
+ \exp(\nu_1 + \lambda_2 - 1) (\nu_1 + \lambda_2 - 1)
+ \nu_1 (1 - \exp(\nu_1 + \lambda_1 - 1) - \exp(\nu_1 + \lambda_2 - 1))
- \lambda_1 \exp(\nu_1 + \lambda_1 - 1) - \lambda_2 \exp(\nu_1 + \lambda_2 - 1)
\]
Example

\[
\min_{x \in \mathbb{R}^3} x_1 \log x_1 + x_2 \log x_2
\]

subject to:

\[
\begin{align*}
1 - x_1 - x_2 &= 0 \\
-x_1 &\leq 0 \\
-x_2 &\leq 0
\end{align*}
\]

\[
L(x_1, x_2, \nu_1, \lambda_1, \lambda_2)
= x_1 \log x_1 + x_2 \log x_2 + \nu_1 \cdot (1 - x_1 - x_2) - \lambda_1 x_1 - \lambda_2 x_2
\]

\[
g(\nu_1, \lambda_1, \lambda_2) = -\exp(\nu_1 + \lambda_1 - 1) - \exp(\nu_1 + \lambda_2 - 1) + \nu_1
\]
The Primal Problem

\[ \min_{x \in \mathbb{R}^n} f_0(x) \]

subject to:

\[ f_i(x) \leq 0, \quad i = 1, \ldots, m \]
\[ h_i(x) = 0, \quad i = 1, \ldots, p \]

Equivalently,

\[ \inf_{x} \sup_{\lambda \geq 0, \nu} L(x, \lambda, \nu) \]

Why are these equivalent?
The Primal Problem

\[
\begin{align*}
\min_{x \in \mathbb{R}^n} f_0(x) \\
\text{subject to:} \\
f_i(x) &\leq 0, \quad i = 1, \ldots, m \\
h_i(x) &= 0, \quad i = 1, \ldots, p
\end{align*}
\]

Equivalently,

\[
\inf \sup_{x, \lambda \geq 0, \nu} L(x, \lambda, \nu) \\
\sup_{\lambda \geq 0, \nu} \left[ f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right] = \infty
\]

whenever \( x \) violates the constraints
The Dual Problem

\[
\sup_{\lambda \geq 0, \nu} g(\lambda, \nu)
\]

Equivalently,

\[
\sup_{\lambda \geq 0, \nu} \inf_x L(x, \lambda, \nu)
\]

- The dual problem is always concave, even if the primal problem is not convex
  - For each \( x \), \( L(x, \lambda, \nu) \) is a linear function in \( \lambda \) and \( \nu \)
  - Maximum (or supremum) of concave functions is concave!

16
Primal vs. Dual

\[
\sup_{\lambda \geq 0, \nu} \inf_x L(x, \lambda, \nu) \leq \inf_x \sup_{\lambda \geq 0, \nu} L(x, \lambda, \nu)
\]

Why?

- \( g(\lambda, \nu) \leq L(x, \lambda, \nu) \) for all \( x \)
- \( L(x', \lambda, \nu) \leq f_0(x') \) for any feasible \( x', \lambda \geq 0 \)
  - \( x \) is feasible if it satisfies all of the constraints
- Let \( x^* \) be the optimal solution to the primal problem and \( \lambda \geq 0 \)
  \[
g(\lambda, \nu) \leq L(x^*, \lambda, \nu) \leq f_0(x^*)
\]
Example

\[
\min_{x \in \mathbb{R}^3} x_1 \log x_1 + x_2 \log x_2
\]

subject to:

\[
1 - x_1 - x_2 = 0 \\
-x_1 \leq 0 \\
-x_2 \leq 0
\]

\[
L(x_1, x_2, \nu_1, \lambda_1, \lambda_2) = x_1 \log x_1 + x_2 \log x_2 + \nu_1 \cdot (1 - x_1 - x_2) - \lambda_1 x_1 - \lambda_2 x_2
\]

\[
g(\nu_1, \lambda_1, \lambda_2) = -\exp(\nu_1 + \lambda_1 - 1) - \exp(\nu_1 + \lambda_2 - 1) + \nu_1
\]

\[
\frac{\partial g}{\partial \nu_1} = -\exp(\nu_1 + \lambda_1 - 1) - \exp(\nu_1 + \lambda_2 - 1) + 1 = 0
\]

\[
g \text{ is a decreasing function of } \lambda_1 \text{ and } \lambda_2, \\
\text{so the optimum is achieved at the boundary } \lambda_1 = \lambda_2 = 0
\]
Example

\[ \min_{x \in \mathbb{R}^3} x_1 \log x_1 + x_2 \log x_2 \]

subject to:

\[ 1 - x_1 - x_2 = 0 \]
\[ -x_1 \leq 0 \]
\[ -x_2 \leq 0 \]

\[ L(x_1, x_2, \nu_1, \lambda_1, \lambda_2) = x_1 \log x_1 + x_2 \log x_2 + \nu_1 \cdot (1 - x_1 - x_2) - \lambda_1 x_1 - \lambda_2 x_2 \]

\[ g(\nu_1, \lambda_1, \lambda_2) = -\exp(\nu_1 + \lambda_1 - 1) - \exp(\nu_1 + \lambda_2 - 1) + \nu_1 \]

\[ \frac{\partial g}{\partial \nu_1} = -\exp(\nu_1 + \lambda_1 - 1) - \exp(\nu_1 + \lambda_2 - 1) + 1 = 0 \]
\[ -\exp(\nu_1 - 1) - \exp(\nu_1 - 1) + 1 = 0 \]
\[ \exp(\nu_1 - 1) = .5 \]
\[ \nu_1 = \log(.5) + 1 \]
• Minimize $x^2 + y^2$ subject to $x + y \geq 1$

• Given a point $z \in \mathbb{R}^n$ and a hyperplane $w^T x + b = 0$, find the projection of the point $z$ onto the hyperplane
Duality

• Under certain conditions, the two optimization problems are equivalent

\[
\sup_{\lambda \geq 0, \nu} \inf_x L(x, \lambda, \nu) = \inf_x \sup_{\lambda \geq 0, \nu} L(x, \lambda, \nu)
\]

• This is called strong duality

• If the inequality is strict, then we say that there is a duality gap

• Size of gap measured by the difference between the two sides of the inequality
Slater’s Condition

For any optimization problem of the form

$$\min_{x \in \mathbb{R}^n} f_0(x)$$

subject to:

$$f_i(x) \leq 0, \quad i = 1, \ldots, m$$
$$Ax = b$$

where $$f_0, \ldots, f_m$$ are convex functions, strong duality holds if there exists an $$x$$ such that

$$f_i(x) < 0, \quad i = 1, \ldots, m$$
$$Ax = b$$
Dual SVM

\[ \min_w \frac{1}{2} \|w\|^2 \]

such that

\[ y_i (w^T x^{(i)} + b) \geq 1, \text{ for all } i \]

- Note that Slater’s condition holds as long as the data is linearly separable
Dual SVM

\[ L(w, b, \lambda) = \frac{1}{2} w^T w + \sum_{i} \lambda_i (1 - y_i (w^T x^{(i)} + b)) \]

Convex in \( w \), so take derivatives to form the dual

\[ \frac{\partial L}{\partial w_k} = w_k + \sum_{i} -\lambda_i y_i x_k^{(i)} = 0 \]

\[ \frac{\partial L}{\partial b} = \sum_{i} -\lambda_i y_i = 0 \]
Dual SVM

\[ L(w, b, \lambda) = \frac{1}{2} w^T w + \sum \lambda_i (1 - y_i (w^T x^{(i)} + b)) \]

Convex in \( w \), so take derivatives to form the dual

\[ w = \sum \lambda_i y_i x^{(i)} \]

\[ \sum \lambda_i y_i = 0 \]
Dual SVM

\[
\max_{\lambda \geq 0} -\frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j x^{(i)T} x^{(j)} + \sum_i \lambda_i
\]

such that

\[
\sum_i \lambda_i y_i = 0
\]

• By strong duality, solving this problem is equivalent to solving the primal problem

• Given the optimal \( \lambda \), we can easily construct \( w \) (\( b \) can be found by complementary slackness...)
Complementary Slackness

• Suppose that there is zero duality gap

• Let $x^*$ be an optimum of the primal and $(\lambda^*, \nu^*)$ be an optimum of the dual

$$f_0(x^*) = g(\lambda^*, \nu^*)$$

$$= \inf_x \left[ f_0(x) + \sum_{i=1}^{m} \lambda_i^* f_i(x) + \sum_{i=1}^{p} \nu_i^* h_i(x) \right]$$

$$\leq f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*) + \sum_{i=1}^{p} \nu_i^* h_i(x^*)$$

$$= f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*)$$

$$\leq f_0(x^*)$$
Complementary Slackness

• This means that

\[ \sum_{i=1}^{m} \lambda_i^* f_i(x^*) = 0 \]

• As \( \lambda \geq 0 \) and \( f_i(x_i^*) \leq 0 \), this can only happen if \( \lambda_i^* f_i(x^*) = 0 \) for all \( i \)

• Put another way,
  
  • If \( f_i(x^*) < 0 \) (i.e., the constraint is not tight), then \( \lambda_i^* = 0 \)
  
  • If \( \lambda_i^* > 0 \), then \( f_i(x^*) = 0 \)

• ONLY applies when there is no duality gap
Dual SVM

\[
\max_{\lambda \geq 0} - \frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j x^{(i)T} x^{(j)} + \sum_i \lambda_i
\]

such that

\[
\sum_i \lambda_i y_i = 0
\]

• By complementary slackness, \( \lambda_i^* > 0 \) means that \( x^{(i)} \) is a support vector (can then solve for \( b \) using \( w \))
Dual SVM

$$\max_{\lambda \geq 0} -\frac{1}{2} \sum_{i} \sum_{j} \lambda_i \lambda_j y_i y_j x^{(i)T} x^{(j)} + \sum_{i} \lambda_i$$

such that

$$\sum_{i} \lambda_i y_i = 0$$

• Takes $O(n^2)$ time just to evaluate the objective function
  • Active area of research to try to speed this up
Dual SVM

\[
\max_{\lambda \geq 0} - \frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j x^{(i)T} x^{(j)} + \sum_i \lambda_i
\]

such that

\[
\sum_i \lambda_i y_i = 0
\]

• The dual formulation only depends on inner products between the data points

• Same thing is true if we use feature vectors instead
Dual SVM

\[
\max_{\lambda \geq 0} -\frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j \Phi(x^{(i)})^T \Phi(x^{(j)}) + \sum_i \lambda_i
\]

such that

\[
\sum_i \lambda_i y_i = 0
\]

• The dual formulation only depends on inner products between the data points

• Same thing is true if we use feature vectors instead
The Kernel Trick

- For some feature vectors, we can compute the inner products quickly, even if the feature vectors are very large

- This is best illustrated by example

Let $\phi(x_1, x_2) = \begin{bmatrix} x_1 x_2 \\ x_2 x_1 \\ x_1^2 \\ x_2^2 \end{bmatrix}$

- $\phi(x_1, x_2)^T \phi(z_1, z_2) = x_1^2 z_1^2 + 2x_1 x_2 z_1 z_2 + x_2^2 z_2^2$

  $= (x_1 z_1 + x_2 z_2)^2$

  $= (x^T z)^2$
The Kernel Trick

• For some feature vectors, we can compute the inner products quickly, even if the feature vectors are very large

• This is best illustrated by example

Let \( \phi(x_1, x_2) = \begin{bmatrix} x_1 x_2 \\ x_2 x_1 \\ x_1^2 \\ x_2^2 \end{bmatrix} \)

\[
\phi(x_1, x_2)^T \phi(z_1, z_2) = x_1^2 z_1^2 + 2x_1 x_2 z_1 z_2 + x_2^2 z_2^2 \\
= (x_1 z_1 + x_2 z_2)^2 \\
= (x^T z)^2
\]

Reduces to a dot product in the original space
The Kernel Trick

• The same idea can be applied for the feature vector $\phi$ of all polynomials of degree (exactly) $d$
  
  • $\phi(x)^T \phi(z) = (x^T z)^d$

• More generally, a kernel is a function $k(x, z) = \phi(x)^T \phi(z)$ for some feature map $\phi$

• Rewrite the dual objective

$$\max_{\lambda \geq 0, \sum_i \lambda_i y_i=0} \left( -\frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j k(x^{(i)}, x^{(j)}) + \sum_i \lambda_i \right)$$
Examples of Kernels

• Polynomial kernel of degree exactly $d$
  
  - $k(x, z) = (x^T z)^d$

• General polynomial kernel of degree $d$ for some $c$
  
  - $k(x, z) = (x^T z + c)^d$

• Gaussian kernel for some $\sigma$
  
  - $k(x, z) = \exp\left(\frac{-\|x - z\|^2}{2\sigma^2}\right)$
    
    - The corresponding $\phi$ is infinite dimensional!

• So many more...
Gaussian Kernels

• Consider the Gaussian kernel

\[
\exp \left( \frac{-(x - z)^T(x - z)}{2\sigma^2} \right) = \exp \left( -\frac{\|x - z\|^2}{2\sigma^2} \right)
\]

\[
= \exp \left( -\frac{\|x\|^2 + 2x^Tz - \|z\|^2}{2\sigma^2} \right)
\]

\[
= \exp \left( -\frac{\|x\|^2}{2\sigma^2} \right) \exp \left( -\frac{\|z\|^2}{2\sigma^2} \right) \exp \left( \frac{x^Tz}{\sigma^2} \right)
\]

• Use the Taylor expansion for \( \exp() \)

\[
\exp \left( \frac{x^Tz}{\sigma^2} \right) = \sum_{n=0}^{\infty} \frac{(x^Tz)^n}{\sigma^{2n}n!}
\]
Gaussian Kernels

• Consider the Gaussian kernel

\[
\exp \left( -\frac{\|x - z\|^2}{2\sigma^2} \right) = \exp \left( -\frac{(x - z)^T (x - z)}{2\sigma^2} \right)
\]

\[
= \exp \left( -\frac{\|x\|^2 + 2x^T z - \|z\|^2}{2\sigma^2} \right)
\]

\[
= \exp \left( -\frac{\|x\|^2}{2\sigma^2} \right) \exp \left( -\frac{\|z\|^2}{2\sigma^2} \right) \exp \left( \frac{x^T z}{\sigma^2} \right)
\]

• Use the Taylor expansion for \( \exp() \)

\[
\exp \left( \frac{x^T z}{\sigma^2} \right) = \sum_{n=0}^{\infty} \frac{(x^T z)^n}{\sigma^{2n} n!}
\]

Polynomial kernels of every degree!
Kernels

• Bigger feature space increases the possibility of overfitting
  • Large margin solutions may still generalize reasonably well

• Alternative: add “penalties” to the objective to disincentivize complicated solutions

\[
\min_w \frac{1}{2} \|w\|^2 + c \cdot (\# \text{ of misclassifications})
\]

• Not a quadratic program anymore (in fact, it’s NP-hard)

• Similar problem to counting the number of misclassifications, no notion of how badly the data is misclassified