Representer Theorem

He He

Slides based on Lecture 5a from David Rosenberg's course material.

CDS, NYU

March 2, 2021

SVM solution is in the "span of the data"

- We found the SVM dual problem can be written as:

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

- Given dual solution α^*, primal solution is $w^* = \sum_{i=1}^{n} \underbrace{\alpha_i^* y_i} x_i$.
- Notice: w^* is a linear combination of training inputs $x_1, \dots, x_n$.
- We refer to this phenomenon by saying "w^* is in the **span of the data**."
    - Or in math, $w^* \in \operatorname{span}(x_1, \dots, x_n)$.

Ridge regression solution is in the "span of the data"

- The ridge regression solution for regularization parameter $\lambda > 0$ is

$$w^* = \underset{w \in \mathbb{R}^d}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left\{ w^T x_i - y_i \right\}^2 + \lambda \|w\|_2^2.$$

- This has a closed form solution (Homework #3):

$$w^* = \left(X^T X + \lambda I \right)^{-1} X^T y,$$

where X is the design matrix, with $x_1, \ldots, x_n$ as rows.

Ridge regression solution is in the "span of the data"

- Rearranging $w^* = \left(X^T X + \lambda I\right)^{-1} X^T y$, we can show that (also Homework #3):

$$
\begin{aligned}
w^* &= X^T \underbrace{\left(\frac{1}{\lambda} y - \frac{1}{\lambda} X w^*\right)}_{\alpha^*} \\
&= X^T \alpha^* = \sum_{i=1}^{n} \alpha_i^* x_i .
\end{aligned}
$$

- So w^* is in the span of the data.
  - i.e. $w^* \in \operatorname{span}(x_1, \ldots, x_n)$

If solution is in the span of the data, we can reparameterize

- The ridge regression solution for regularization parameter $\lambda > 0$ is

$$w^* = \underset{w \in \mathbb{R}^d}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left\{ w^T x_i - y_i \right\}^2 + \lambda \|w\|_2^2.$$

- We now know that $w^* \in \operatorname{span}(x_1, \ldots, x_n) \subset \mathbb{R}^d$.

- So rather than minimizing over all of $\mathbb{R}^d$, we can minimize over $\operatorname{span}(x_1, \ldots, x_n)$.

$$w^* = \underset{w \in \operatorname{span}(x_1, \ldots, x_n)}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left\{ w^T x_i - y_i \right\}^2 + \lambda \|w\|_2^2.$$

- Let's reparameterize the objective by replacing w as a linear combination of the inputs.

If solution is in the span of the data, we can reparameterize

- Note that for any $w \in \text{span}(x_1, \ldots, x_n)$, we have $w = X^T\alpha$, for some $\alpha \in \mathbb{R}^n$.
- So let's replace w with $X^T\alpha$ in our optimization problem:

$$
\begin{aligned}
\text{[original] } w^* &= \underset{w \in \mathbb{R}^d}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left\{ w^T x_i - y_i \right\}^2 + \lambda \|w\|_2^2 \\
\text{[reparameterized] } \alpha^* &= \underset{\alpha \in \mathbb{R}^n}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left\{ \left(X^T\alpha\right)^T x_i - y_i \right\}^2 + \lambda \|X^T\alpha\|_2^2.
\end{aligned}
$$

- To get w^* from the reparameterized optimization problem, we just take $w^* = X^T\alpha^*$.
- We changed the dimension of our optimization variable from d to n. Is this useful?

Consider very large feature spaces

- Suppose we have a 300-million dimension feature space [very large]
  - (e.g. using high order monomial interaction terms as features, as described last lecture)
- Suppose we have a training set of 300,000 examples [fairly large]
- In the original formulation, we solve a 300-million dimension optimization problem.
- In the reparameterized formulation, we solve a 300,000-dimension optimization problem.
- This is why we care about when the solution is in the span of the data.
- This reparameterization is interesting when we have more features than data ($d \gg n$).

What's next?

- For SVM and ridge regression, we found that the solution is in the span of the data.
    - derived in two rather ad-hoc ways
- Up next: The Representer Theorem, which shows that this "span of the data" result occurs far more generally, and we prove it using basic linear algebra.

Math Review: Inner Product Spaces and Hilbert Spaces

Hypothesis spaces we've seen so far

Finite-dimensional vector space (linear functions):

$$\mathcal{H} = \left\{ f : \mathcal{X} \to \mathrm{R} \mid f(x) = w^T x, \quad w, x \in \mathrm{R}^d \right\} .$$

To consider more complex input spaces (e.g. text, images), we use a feature map $\phi : \mathcal{X} \to \mathcal{F}$:

$$\mathcal{H} = \left\{ f : \mathcal{X} \to \mathrm{R} \mid f(x) = w^T \phi(x) \right\} .$$

- ϕ does not have to be linear.
- The feature space $\mathcal{F}$ can be R^d (Euclidean space) or an infinite-dimensional vector space.
- We would like more structure on $\mathcal{F}$.

Inner Product Space (or "Pre-Hilbert" Spaces)

An **inner product space** (over reals) is a vector space $\mathcal{V}$ with an **inner product**, which is a mapping

$$\langle \cdot, \cdot \rangle : \mathcal{V} \times \mathcal{V} \to \mathbf{R}$$

that has the following properties: $\forall x, y, z \in \mathcal{V}$ and $a, b \in \mathbf{R}$:

- Symmetry: $\langle x, y \rangle = \langle y, x \rangle$
- Linearity: $\langle ax + by, z \rangle = a \langle x, z \rangle + b \langle y, z \rangle$ $\begin{cases} \alpha \langle x, y \rangle = \langle \alpha x, y \rangle \\ \langle x + z, y \rangle = \langle x, y \rangle + \langle z, y \rangle \end{cases}$
- Positive-definiteness: $\langle x, x \rangle \geqslant 0$ and $\langle x, x \rangle = 0 \iff x = 0_{\mathcal{V}}$.

To show a function $\langle \cdot, \cdot \rangle$ is an inner product, we need to check the above conditions.

Exercise: show that $\langle x, y \rangle \stackrel{\text{def}}{=} x^T y$ is an inner product on $\mathbf{R}^d$.

Norm from Inner Product

Inner product is nice because it gives us notions of "size", "distance", "angle" in the vector space.

For an inner product space, we can ddefine a norm as

$$\|x\| \stackrel{\text{def}}{=} \sqrt{\langle x, x \rangle}.$$

Example

R^d with standard Euclidean inner product is an inner product space:

$$\langle x, y \rangle := x^T y \qquad \forall x, y \in \mathrm{R}^d.$$

Norm is

$$\|x\| = \sqrt{x^T x}.$$

Orthogonality (Definitions)

Definition

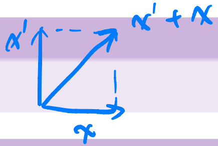
Two vectors are **orthogonal** if $\langle x, x' \rangle = 0$. We denote this by $x \perp x'$.

Definition

x is orthogonal to a set S, i.e. $x \perp S$, if $x \perp s$ for all $x \in S$.

Theorem (Pythagorean Theorem)

If $x \perp x'$, *then* $\|x+x'\|^2 = \|x\|^2 + \|x'\|^2$.



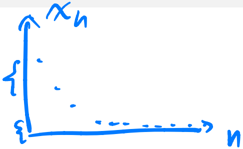
Proof.

We have

$$\begin{aligned} \|x+x'\|^2 &= \langle x+x', x+x'\rangle \quad \text{by def} \\ &= \langle x,x\rangle + \underbrace{\cancel{\langle x,x'\rangle}}_{0} + \underbrace{\cancel{\langle x',x\rangle}}_{0} + \langle x',x'\rangle \quad \text{lin.} \\ &= \|x\|^2 + \|x'\|^2. \quad \text{by orthogonality} \end{aligned}$$

$\square$

Hilbert Space



- A pre-Hilbert space is a vector space equipped with an inner product.
- We need an additional technical condition for Hilbert space: **completeness**.
- A space is **complete** if all Cauchy sequences in the space converge to a point in the space.

Definition

A **Hilbert space** is a complete inner product space.

Example

Any finite dimensional inner produce space is a Hilbert space.

The Representer Theorem

Generalize from SVM Objective

- SVM objective:

$$\min_{w \in \mathbf{R}^d} \frac{1}{2}\|w\|^2 + \frac{c}{n}\sum_{i=1}^{n} \max\left(0, 1 - y_i\left[\langle w, x_i \rangle\right]\right).$$

- **Generalized objective**:

$\phi(x_1)$

$$\min_{w \in \mathcal{H}} R\left(\|w\|\right) + L\left(\langle w, x_1 \rangle, \ldots, \langle w, x_n \rangle\right),$$

where

- $w, x_1, \ldots, x_n \in \mathcal{H}$ for some Hilbert space $\mathcal{H}$. (We typically have $\mathcal{H} = \mathbf{R}^d$.)
- $\|\cdot\|$ is the norm corresponding to the inner product of $\mathcal{H}$. (i.e. $\|w\| = \sqrt{\langle w, w \rangle}$)
- $R : [0, \infty) \to \mathbf{R}$ is nondecreasing (**Regularization term**), and
- $L : \mathbf{R}^n \to \mathbf{R}$ is arbitrary (**Loss term**).

General Objective Function for Linear Hypothesis Space (Details)

- **Generalized objective**:

$$\min_{w \in \mathcal{H}} R\left(\|w\|\right) + L\left(\langle w, x_1 \rangle, \ldots, \langle w, x_n \rangle\right)$$

- We can map x_i to a feature space.
- The prediction/score function $x \mapsto \langle w, x \rangle$ is linear in w.

General Objective Function for Linear Hypothesis Space (Details)

- **Generalized objective**:

$$\min_{w \in \mathcal{H}} R(\|w\|) + L(\langle w, x_1 \rangle, \ldots, \langle w, x_n \rangle)$$

- Ridge regression and SVM are of this form. (Verify this!)
- What if we penalize with $\lambda \|w\|_2$ instead of $\lambda \|w\|_2^2$? Yes!
- What if we use lasso regression? No! ℓ_1 norm does not correspond to an inner product.

The Representer Theorem: Quick Summary

- **Generalized objective**:

$$w^* = \underset{w \in \mathcal{H}}{\arg\min}\, R\left(\|w\|\right) + L\left(\langle w, x_1 \rangle, \ldots, \langle w, x_n \rangle\right)$$

- Representer theorem tells us we can look for w^* in the span of the data:

$$w^* = \underset{w \in \text{span}(x_1, \ldots, x_n)}{\arg\min} R\left(\|w\|\right) + L\left(\langle w, x_1 \rangle, \ldots, \langle w, x_n \rangle\right).$$

- So we can reparameterize as before:

$$\alpha^* = \underset{\alpha \in \mathbf{R}^n}{\arg\min}\, R\left(\left\|\sum_{i=1}^n \alpha_i x_i\right\|\right) + L\left(\left\langle \sum_{i=1}^n \alpha_i x_i, x_1 \right\rangle, \ldots, \left\langle \sum_{i=1}^n \alpha_i x_i, x_n \right\rangle\right).$$

- Our reparameterization trick applies much more broadly than SVM and ridge.

The Representer Theorem

Theorem (Representer Theorem)

Let

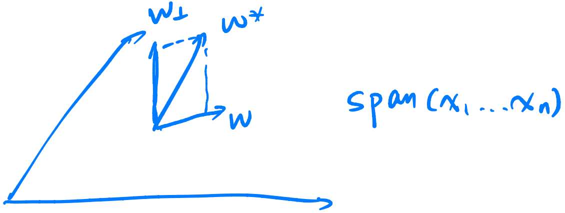
$$J(w) = R\left(\|w\|\right) + L\left(\langle w, x_1\rangle, \ldots, \langle w, x_n\rangle\right),$$

where

- $w, x_1, \ldots, x_n \in \mathcal{H}$ *for some Hilbert space* $\mathcal{H}$. *(We typically have* $\mathcal{H} = \mathrm{R}^d$.*)*
- $\|\cdot\|$ *is the norm corresponding to the inner product of* $\mathcal{H}$. *(i.e.* $\|w\| = \sqrt{\langle w, w\rangle}$*)*
- $R : [0, \infty) \to \mathrm{R}$ *is nondecreasing* ***(Regularization term)****, and*
- $L : \mathrm{R}^n \to \mathrm{R}$ *is arbitrary* ***(Loss term)****.*

Then it ***has a minimizer of the form*** $w^* = \sum_{i=1}^n \alpha_i x_i$.

The Representer Theorem (Proof sketch)



$$\langle w^*, x\rangle = \langle w + w_\perp, x\rangle$$

$$= \langle w, x\rangle + \cancel{\langle w_\perp, x\rangle} \quad 0 \quad \perp$$

prediction score

$$= \langle w, x\rangle$$

$$\|w\| < \|w^*\|$$

$$R(\|w\|) < R(\|w^*\|)$$

Reparameterizing our Generalized Objective Function

Rewriting the Objective Function

- Define the training score function $s : \mathbf{R}^d \to \mathbf{R}^n$ by

$$s(w) = \begin{pmatrix} \langle w, x_1 \rangle \\ \vdots \\ \langle w, x_n \rangle \end{pmatrix},$$

which gives the **training score vector** for any w.

- We can then rewrite the objective function as

$$J(w) = R(\|w\|) + L(s(w)),$$

where now $L : \mathbf{R}^{n \times 1} \to \mathbf{R}$ takes a column vector as input.

- This will allow us to have a slick reparameterized version...

Reparameterize the Generalized Objective

- By the Representer Theorem, it's sufficient to minimize $J(w)$ for w of the form $\sum_{i=1}^{n} \alpha_i x_i$.
- Plugging this form into $J(w)$, we see we can just minimize

$$J_0(\alpha) = R\left(\left\|\sum_{i=1}^{n} \alpha_i x_i\right\|\right) + L\left(s\left(\sum_{i=1}^{n} \alpha_i x_i\right)\right)$$

$\|w\|$ $\quad s(w)$

over $\alpha = (\alpha_1, \ldots, \alpha_n)^T \in \mathrm{R}^{n \times 1}$.

- With some new notation, we can substantially simplify
  - the norm piece $\|w\| = \|\sum_{i=1}^{n} \alpha_i x_i\|$, and
  - the score piece $s(w) = s\left(\sum_{i=1}^{n} \alpha_i x_i\right)$.

Simplifying the Reparameterized Norm

- For the norm piece $\|w\| = \|\sum_{i=1}^{n} \alpha_i x_i\|$, we have

$$
\begin{aligned}
\|w\|^2 &= \langle w, w \rangle && \text{def} \\
&= \left\langle \sum_{i=1}^{n} \alpha_i x_i, \sum_{j=1}^{n} \alpha_j x_j \right\rangle && \text{repr.} \\
&= \sum_{i,j=1}^{n} \alpha_i \alpha_j \underbrace{\langle x_i, x_j \rangle}_{K_{ij}}. && \text{lin.} \quad \alpha^T K \alpha
\end{aligned}
$$

- This expression involves the n^2 inner products between all pairs of input vectors.
- We often put those values together into a matrix (Gram/Kernel matrix).

Example: Gram Matrix for the Dot Product

- Consider $x_1, \ldots, x_n \in \mathbf{R}^{d \times 1}$ with the standard inner product $\langle x, x' \rangle = x^T x'$.
- Let $X \in \mathbf{R}^{n \times d}$ be the **design matrix**, which has each input vector as a row:

$$X = \begin{pmatrix} -x_1^T- \\ \vdots \\ -x_n^T- \end{pmatrix}.$$

- Then the Gram matrix is $K_{ij} = k(x_i, x_j) = \langle x_i, x_j \rangle = x_i^T x_j$

$$\begin{aligned} K &= \begin{pmatrix} x_1^T x_1 & \cdots & x_1^T x_n \\ \vdots & \ddots & \cdots \\ x_n^T x_1 & \cdots & x_n^T x_n \end{pmatrix} = \begin{pmatrix} -x_1^T- \\ \vdots \\ -x_n^T- \end{pmatrix} \begin{pmatrix} | & \cdots & | \\ x_1 & \cdots & x_n \\ | & \cdots & | \end{pmatrix} \\ &= XX^T \end{aligned}$$

Simplifying the Reparametrized Norm

- With $w = \sum_{i=1}^{n} \alpha_i x_i$, we have

$$\begin{aligned}
\|w\|^2 &= \langle w, w \rangle \\
&= \left\langle \sum_{i=1}^{n} \alpha_i x_i, \sum_{j=1}^{n} \alpha_j x_j \right\rangle \\
&= \sum_{i,j=1}^{n} \alpha_i \alpha_j \underbrace{\langle x_i, x_j \rangle}_{K_{ij}} \\
&= \alpha^T K \alpha.
\end{aligned}$$

Simplifying the Training Score Vector

- The score for x_j for $w = \sum_{i=1}^n \alpha_i x_i$ is

$$\langle w, x_j \rangle = \left\langle \sum_{i=1}^n \alpha_i x_i, x_j \right\rangle = \sum_{i=1}^n \alpha_i \langle x_i, x_j \rangle$$

- The training score vector is

$s(w)$

$$\begin{aligned}
= s\left(\sum_{i=1}^n \alpha_i x_i\right) &= \begin{pmatrix} \sum_{i=1}^n \alpha_i \langle x_i, x_1 \rangle \\ \vdots \\ \sum_{i=1}^n \alpha_i \langle x_i, x_n \rangle \end{pmatrix} = \begin{pmatrix} \alpha_1 \langle x_1, x_1 \rangle + \cdots + \alpha_n \langle x_n, x_1 \rangle \\ \vdots \\ \alpha_1 \langle x_1, x_n \rangle + \cdots + \alpha_n \langle x_n, x_n \rangle \end{pmatrix} \\
= \begin{bmatrix} \langle w, x_1 \rangle \\ \vdots \\ \langle w, x_n \rangle \end{bmatrix} &= \begin{pmatrix} \langle x_1, x_1 \rangle & \cdots & \langle x_1, x_n \rangle \\ \vdots & \ddots & \cdots \\ \langle x_n, x_1 \rangle & \cdots & \langle x_n, x_n \rangle \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} \\
&= K\alpha
\end{aligned}$$

Reparameterized Objective

- Putting it all together, our reparameterized objective function can be written as

$$\begin{aligned} J_0(\alpha) &= R\left(\left\|\sum_{i=1}^{n}\alpha_i x_i\right\|\right) + L\left(s\left(\sum_{i=1}^{n}\alpha_i x_i\right)\right) \\ &= R\left(\sqrt{\alpha^T K \alpha}\right) + L(K\alpha), \end{aligned}$$

  which we minimize over $\alpha \in \mathbb{R}^n$.
- **All information** needed about $x_1, \ldots, x_n$ is summarized in the Gram matrix K.
- We're now minimizing over $\mathbb{R}^n$ rather than $\mathbb{R}^d$.
- If $d \gg n$, this can be a big win computationally (at least once K is computed).

Reparameterizing Predictions

- Suppose we've found

$$\alpha^* \in \underset{\alpha \in \mathbf{R}^n}{\arg\min}\, R\left(\sqrt{\alpha^T K \alpha}\right) + L\left(K\alpha\right).$$

- Then we know $w^* = \sum_{i=1}^n \alpha^* x_i$ is a solution to

$$\underset{w \in \mathcal{H}}{\arg\min}\, R\left(\|w\|\right) + L\left(\langle w, x_1\rangle, \ldots, \langle w, x_n\rangle\right).$$

- The prediction on a new point $x \in \mathcal{H}$ is

$= \langle \sum_{i=1}^{n} \alpha_i x_i, x \rangle$

$$\hat{f}(x) = \langle w^*, x\rangle = \sum_{i=1}^n \alpha_i^* \langle x_i, x\rangle.$$

test ex.

i-th train ex

- To make a new prediction, we may need to touch all the training inputs $x_1, \ldots, x_n$.

More Notation

- It will be convenient to define the following column vector for any $x \in \mathcal{H}$:

$$k_x = \begin{pmatrix} \langle x_1, x \rangle \\ \vdots \\ \langle x_n, x \rangle \end{pmatrix}$$

test ex.

- Then we can write our predictions on a new point x as

$$\hat{f}(x) = k_x^T \alpha^*$$

Summary So Far

- Original plan:
  - Find $w^* \in \arg\min_{w \in \mathcal{H}} R\left(\|w\|\right) + L\left(\langle w, x_1 \rangle, \ldots, \langle w, x_n \rangle\right)$
  - Predict with $\hat{f}(x) = \langle w^*, x \rangle$.
- We showed that the following is equivalent:
  - Find $\alpha^* \in \arg\min_{\alpha \in \mathbf{R}^n} R\left(\sqrt{\alpha^T K \alpha}\right) + L\left(K\alpha\right)$
  - Predict with $\hat{f}(x) = k_x^T \alpha^*$, where

$$K = \begin{pmatrix} \langle x_1, x_1 \rangle & \cdots & \langle x_1, x_n \rangle \\ \vdots & \ddots & \ldots \\ \langle x_n, x_1 \rangle & \cdots & \langle x_n, x_n \rangle \end{pmatrix} \quad \text{and} \quad k_x = \begin{pmatrix} \langle x_1, x \rangle \\ \vdots \\ \langle x_n, x \rangle \end{pmatrix}$$

  - Every element $x \in \mathcal{H}$ occurs inside an inner products with a training input $x_i \in \mathcal{H}$.

Kernelization

Definition

A method is **kernelized** if every feature vector $\psi(x)$ only appears inside an inner product with another feature vector $\psi(x')$. This applies to both the optimization problem and the prediction function.

- Here we are using $\psi(x) = x$. Thus finding

$$\alpha^* \in \operatorname*{arg\,min}_{\alpha \in \mathrm{R}^n} R\left(\sqrt{\alpha^T K \alpha}\right) + L\left(K\alpha\right)$$

and making predictions with $\hat{f}(x) = k_x^T \alpha^*$ is a **kernelization** of finding

$$w^* \in \operatorname*{arg\,min}_{w \in \mathcal{H}} R\left(\|w\|\right) + L\left(\langle w, x_1 \rangle, \ldots, \langle w, x_n \rangle\right)$$

and making predictions with $\hat{f}(x) = \langle w^*, x \rangle$.

Summary

- We used duality for SVM and bare hands for ridge regression to find their kernelized version.
- Our principle tool for kernelization is reparameterization by the representer theorem.
- Once kernelized, we can apply the kernel trick: doesn't need to represent $\phi(x)$ explicitly.