Support vector machines

CS 446

2020-12-27 (a50ee91)
Another algorithm for linear prediction. Why?!
“pytorch meta-algorithm”.

5. Tweak 1-4 until training error is small.
6. Tweak 1-5, possibly reducing model complexity, until testing error is small.

Recall: step 5 is easy, step 6 is complicated.
Nice way to reduce complexity: favorable inductive bias.
5. Tweak 1-4 until training error is small.
6. Tweak 1-5, possibly reducing model complexity, until testing error is small.

Recall: step 5 is easy, step 6 is complicated.
Nice way to reduce complexity: favorable inductive bias.

- **Support vector machines (SVMs)** popularized a very influential inductive bias: maximum margin predictors.
  This concept has consequence beyond linear predictors and beyond supervised learning.

- **Nonlinear SVMs**, via kernels, are also highly influential.
  E.g., we will revisit them in the deep network lectures.
Plan for SVM

- Hard-margin SVM.
- Soft-margin SVM.
- SVM duality.
- Nonlinear SVM: kernels
Linearly separable data means there exists $u \in \mathbb{R}^d$ which perfectly classifies all training points:

$$\min_i y_i x_i^T u > 0.$$
**Linearly separable data** means there exists $u \in \mathbb{R}^d$ which perfectly classifies all training points:

$$\min_i y_i x_i^T u > 0.$$  

We have many ways to solve for $u$:

- Logistic regression.
- Perceptron.
- Convex programming (convex function subject to convex set constraint).
Maximum margin solution

Best linear classifier on population

Why use the maximum margin solution?

(i) Uniquely determined by $S$, unlike the linear program.
(ii) It is a particular inductive bias—i.e., an assumption about the problem—that seems to be commonly useful.

▶ We've seen inductive bias: least squares and logistic regression choose different predictors on same data.

▶ This particular bias (margin maximization) is common in machine learning, has many nice properties.

Key insight: can express this as another convex program.
Maximum margin solution

Best linear classifier on population

Arbitrary linear separator on training data $S$

Why use the maximum margin solution?

(i) Uniquely determined by $S$, unlike the linear program.

(ii) It is a particular inductive bias—i.e., an assumption about the problem—that seems to be commonly useful.

We've seen inductive bias: least squares and logistic regression choose different predictors on same data.

This particular bias (margin maximization) is common in machine learning, has many nice properties.

Key insight: can express this as another convex program.
Maximum margin solution

Best linear classifier on population

Arbitrary linear separator on training data $S$

Why use the maximum margin solution?

(i) Uniquely determined by $S$, unlike the linear program.

(ii) It is a particular inductive bias—i.e., an assumption about the problem—that seems to be commonly useful.

▶ We've seen inductive bias: least squares and logistic regression choose different predictors on same data.

▶ This particular bias (margin maximization) is common in machine learning, has many nice properties.

Key insight: can express this as another convex program.
Maximum margin solution

- Best linear classifier on population
- Arbitrary linear separator on training data $S$
- Maximum margin solution on training data $S$

Why use the maximum margin solution?

(i) Uniquely determined by $S$, unlike the linear program.

(ii) It is a particular inductive bias—i.e., an assumption about the problem—that seems to be commonly useful.

▶ We've seen inductive bias: least squares and logistic regression choose different predictors on same data.

▶ This particular bias (margin maximization) is common in machine learning, has many nice properties.

Key insight: can express this as another convex program.
Maximum margin solution

Best linear classifier on population

Arbitrary linear separator on training data $S$

Maximum margin solution on training data $S$

Why use the maximum margin solution?

(i) Uniquely determined by $S$, unlike the linear program.

(ii) It is a particular inductive bias—i.e., an assumption about the problem—that seems to be commonly useful.

We've seen inductive bias: least squares and logistic regression choose different predictors on same data.

This particular bias (margin maximization) is common in machine learning, has many nice properties.
Why use the maximum margin solution?
(i) Uniquely determined by $S$, unlike the linear program.
(ii) It is a particular inductive bias—i.e., an assumption about the problem—that seems to be commonly useful.

- We've seen inductive bias:
  least squares and logistic regression choose different predictors on same data.
- This particular bias (margin maximization) is common in machine learning, has many nice properties.
Maximum margin solution

Best linear classifier on population
Arbitrary linear separator on training data \( S \)
Maximum margin solution on training data \( S \)

**Why use the maximum margin solution?**
(i) Uniquely determined by \( S \), unlike the linear program.
(ii) It is a particular inductive bias—i.e., an assumption about the problem—that seems to be commonly useful.

- We've seen inductive bias:
  least squares and logistic regression choose different predictors on same data.

- This particular bias (margin maximization) is common in machine learning, has many nice properties.

**Key insight:** can express this as another convex program.
Distance to decision boundary

Suppose $w \in \mathbb{R}^d$ satisfies $\min_{(x, y) \in S} yx^T w > 0$. 

"Maximum margin" shouldn't care about scaling; $w$ and $10w$ should be equally good.

Thus for each direction $w/\|w\|$, we can fix a scaling. Let $(\tilde{x}, \tilde{y})$ be any example in $S$ that achieves the minimum.

Rescale $w$ so that $\tilde{y} \tilde{x}^T w = 1$.

(Now scaling is fixed.) Distance from $\tilde{y} \tilde{x}$ to $H$ is $\tilde{y} \tilde{x}^T w / \|w\| = 1 / \|w\|$. This is the (normalized minimum) margin.

This gives optimization problem $\max 1 / \|w\|$ subj. to $\min_{(x, y) \in S} yx^T w = 1$.

Refinements: (a) can make constraint $\forall i, y_i x_i^T w \geq 1$; (b) can replace $\max 1 / \|w\|$ with $\min \|w\|^2$. 

6 / 37
Suppose $w \in \mathbb{R}^d$ satisfies $\min_{(x, y) \in S} yx^Tw > 0$. 

- “Maximum margin” shouldn’t care about scaling; $w$ and $10w$ should be equally good.
- Thus for each direction $w/\|w\|$, we can fix a scaling.
Distance to decision boundary

Suppose $w \in \mathbb{R}^d$ satisfies $\min_{(x,y) \in S} yx^T w > 0$.

- "Maximum margin" shouldn't care about scaling; $w$ and $10w$ should be equally good.
- Thus for each direction $w/\|w\|$, we can fix a scaling.

Let $(\tilde{x}, \tilde{y})$ be any example in $S$ that achieves the minimum.
Distance to decision boundary

Suppose \( w \in \mathbb{R}^d \) satisfies \( \min_{(x,y) \in S} yx^T w > 0 \).

- “Maximum margin” shouldn’t care about scaling; \( w \) and \( 10w \) should be equally good.
- Thus for each direction \( w/\|w\| \), we can fix a scaling.

Let \((\tilde{x}, \tilde{y})\) be any example in \( S \) that achieves the minimum.

Refinements: (a) can make constraint \( \forall i, y_i x_i^T w \geq 1 \); (b) can replace \( \max 1/\|w\| \) with \( \min \|w\|_2 \).
Distance to decision boundary

Suppose $w \in \mathbb{R}^d$ satisfies $\min_{(x, y) \in S} yx^T w > 0$.

- “Maximum margin” shouldn’t care about scaling; $w$ and $10w$ should be equally good.
- Thus for each direction $w/\|w\|$, we can fix a scaling.

Let $(\tilde{x}, \tilde{y})$ be any example in $S$ that achieves the minimum.

- Rescale $w$ so that $\tilde{y}\tilde{x}^T w = 1$.
  (Now scaling is fixed.)
Distance to decision boundary

Suppose \( w \in \mathbb{R}^d \) satisfies \( \min_{(x, y) \in S} yx^T w > 0 \).

- “Maximum margin” shouldn’t care about scaling; \( w \) and \( 10w \) should be equally good.
- Thus for each direction \( \frac{w}{\|w\|} \), we can fix a scaling.

Let \((\tilde{x}, \tilde{y})\) be any example in \( S \) that achieves the minimum.

- Rescale \( w \) so that \( \tilde{y}\tilde{x}^T w = 1 \).
  (Now scaling is fixed.)

- Distance from \( \tilde{y}\tilde{x} \) to \( H \) is \( \frac{\tilde{y}\tilde{x}^T w}{\|w\|} = \frac{1}{\|w\|} \).
  This is the (normalized minimum) margin.
Distance to decision boundary

Suppose $w \in \mathbb{R}^d$ satisfies $\min_{(x,y) \in S} yx^T w > 0$.

- “Maximum margin” shouldn’t care about scaling; $w$ and $10w$ should be equally good.
- Thus for each direction $w/\|w\|$, we can fix a scaling.

Let $(\tilde{x}, \tilde{y})$ be any example in $S$ that achieves the minimum.

- Rescale $w$ so that $\tilde{y}\tilde{x}^T w = 1$.
  (Now scaling is fixed.)

- Distance from $\tilde{y}\tilde{x}$ to $H$ is $\frac{\tilde{y}\tilde{x}^T w}{\|w\|} = \frac{1}{\|w\|}$.
  This is the (normalized minimum) margin.

- This gives optimization problem
  \[
  \max \frac{1}{\|w\|} \quad \text{subj. to } \min_{(x,y) \in S} yx^T w = 1. 
  \]

Refinements: (a) can make constraint $\forall i, y_i x_i^T w \geq 1$; (b) can replace
max $1/\|w\|$ with $\min \|w\|^2$. 
Maximum margin linear classifier

The solution $\hat{w}$ to the following mathematical optimization problem:

$$
\min_{w \in \mathbb{R}^d} \quad \frac{1}{2} \|w\|_2^2 \\
\text{s.t.} \quad yx^T w \geq 1 \quad \text{for all } (x, y) \in S
$$

gives the linear classifier with the largest minimum margin on $S$—i.e., the maximum margin linear classifier or support vector machine (SVM) classifier.
Maximum margin linear classifier

The solution $\hat{w}$ to the following mathematical optimization problem:

$$\min_{w \in \mathbb{R}^d} \quad \frac{1}{2} \|w\|^2_2$$

subject to:

$$y x^T w \geq 1 \quad \text{for all } (x, y) \in S$$

gives the linear classifier with the largest minimum margin on $S$—i.e., the maximum margin linear classifier or support vector machine (SVM) classifier.

- This is a convex optimization problem: minimization of a convex function, subject to a convex constraint.
- There are many solvers for this problem; it is an area of active research.
- The feasible set is nonempty when $S$ is linearly separable; in this case, the solution is unique.
- Note: some presentations explicitly include and encode the appended “1” feature on $x$; we will not.
The convex program made no sense if the data is not linearly separable. It is sometimes called the hard-margin SVM.
Now we develop a soft-margin SVM for non-separable data.
Soft-margin SVMs (Cortes and Vapnik, 1995)

Start from the hard-margin program:

\[
\begin{align*}
\min_{w \in \mathbb{R}^d} & \quad \frac{1}{2} \|w\|_2^2 \\
\text{s.t.} & \quad y_i x_i^T w \geq 1 \quad \text{for all } i = 1, 2, \ldots, n.
\end{align*}
\]

Suppose it is infeasible (no linear separators).
Soft-margin SVMs (Cortes and Vapnik, 1995)

Start from the hard-margin program:

\[
\begin{align*}
\min_{\textstyle w \in \mathbb{R}^d} & \quad \frac{1}{2} \left\| w \right\|_2^2 \\
\text{s.t.} & \quad y_i x_i^T w \geq 1 \quad \text{for all } i = 1, 2, \ldots, n.
\end{align*}
\]

Suppose it is infeasible (no linear separators).

Introduce slack variables \( \xi_1, \ldots, \xi_n \geq 0 \), and a trade-off parameter \( C > 0 \):

\[
\begin{align*}
\min_{\textstyle w \in \mathbb{R}^d, \xi_1, \ldots, \xi_n \in \mathbb{R}} & \quad \frac{1}{2} \left\| w \right\|_2^2 + C \sum_{i=1}^{n} \xi_i \\
\text{s.t.} & \quad y_i x_i^T w \geq 1 - \xi_i \quad \text{for all } i = 1, 2, \ldots, n, \\
& \quad \xi_i \geq 0 \quad \text{for all } i = 1, 2, \ldots, n,
\end{align*}
\]

which is always feasible. This is called soft-margin SVM.

(Slack variables are auxiliary variables; not needed to form the linear classifier.)
Interpretation of slack variables

\[
\begin{align*}
\min_{w \in \mathbb{R}^d, \xi_1, \ldots, \xi_n \in \mathbb{R}} & \quad \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i \\
\text{s.t.} & \quad y_i x_i^T w \geq 1 - \xi_i & \text{for all } i = 1, 2, \ldots, n, \\
& \quad \xi_i \geq 0 & \text{for all } i = 1, 2, \ldots, n.
\end{align*}
\]

For given \( w \), \( \xi_i/\|w\|_2 \) is distance that \( x_i \) would have to move to satisfy

\[
y_i x_i^T w \geq 1.
\]
Another interpretation of slack variables

Constraints with non-negative slack variables:

\[
\begin{align*}
\min_{\mathbf{w} \in \mathbb{R}^d, \xi_1, \ldots, \xi_n \in \mathbb{R}} & \quad \frac{1}{2} \| \mathbf{w} \|_2^2 + C \sum_{i=1}^{n} \xi_i \\
\text{s.t.} & \quad y_i \mathbf{x}_i^T \mathbf{w} \geq 1 - \xi_i \quad \text{for all } i = 1, 2, \ldots, n, \\
& \quad \xi_i \geq 0 \quad \text{for all } i = 1, 2, \ldots, n.
\end{align*}
\]
Another interpretation of slack variables

Constraints with non-negative slack variables:

\[
\min_{\mathbf{w} \in \mathbb{R}^d, \xi_1, \ldots, \xi_n \in \mathbb{R}} \quad \frac{1}{2} \| \mathbf{w} \|_2^2 + C \sum_{i=1}^{n} \xi_i \\
\text{s.t.} \quad y_i \mathbf{x}_i^T \mathbf{w} \geq 1 - \xi_i \quad \text{for all } i = 1, 2, \ldots, n, \\
\quad \xi_i \geq 0 \quad \text{for all } i = 1, 2, \ldots, n.
\]

Equivalent unconstrained form:
Given \( \mathbf{w} \), the optimal \( \xi_i \) is \( \max\{0, 1 - y_i \mathbf{x}_i^T \mathbf{w}\} \), thus

\[
\min_{\mathbf{w} \in \mathbb{R}^d} \quad \frac{1}{2} \| \mathbf{w} \|_2^2 + C \sum_{i=1}^{n} \left[ 1 - y_i \mathbf{x}_i^T \mathbf{w} \right]_+.
\]

Notation: \( [a]_+ := \max\{0, a\} \) (ReLU!).
Another interpretation of slack variables

Constraints with non-negative slack variables:

\[
\min_{w \in \mathbb{R}^d, \xi_1, \ldots, \xi_n \in \mathbb{R}} \quad \frac{1}{2} \|w\|^2_2 + C \sum_{i=1}^{n} \xi_i \\
\text{s.t.} \quad y_i x_i^T w \geq 1 - \xi_i \quad \text{for all } i = 1, 2, \ldots, n, \\
\xi_i \geq 0 \quad \text{for all } i = 1, 2, \ldots, n.
\]

Equivalent unconstrained form:
Given \( w \), the optimal \( \xi_i \) is \( \max\{0, 1 - y_i x_i^T w\} \), thus

\[
\min_{w \in \mathbb{R}^d} \quad \frac{1}{2} \|w\|^2_2 + C \sum_{i=1}^{n} \left[1 - y_i x_i^T w\right]_+.
\]

Notation: \([a]_+ := \max\{0, a\} \) (ReLU!).

\([1 - yx^T w]_+ \) is hinge loss of \( w \) on example \((x, y)\).
Unconstrained soft-margin SVM

This form is a regularized ERM:

$$\min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{2}\|\mathbf{w}\|^2_2 + C \sum_{i=1}^{n} \ell_{\text{hinge}}(y_i \mathbf{x}_i^T \mathbf{w})$$

where $\ell_{\text{hinge}}(z) = \max\{0, 1 - z\}$.

- We can equivalently substitute $C = 1$ and write $\frac{\lambda}{2}\|\mathbf{w}\|^2$.
- $C$ is a hyper-parameter, with no good search procedure.
okay this is as far as I got.
maybe just do one of the dual derivations and move other to appendix.
A convex program is an optimization problem (minimization or maximization) where a convex objective is minimized over a convex constraint (feasible) set.
A **convex program** is an optimization problem (minimization or maximization) where a convex objective is minimized over a convex constraint (feasible) set.

Every convex program has a corresponding **dual program**. There is a rich theory about this correspondence. For the SVM, the dual has many nice properties:

- Clarifies the role of **support vectors**.
- Leads to a nice nonlinear approach via **kernels**.
- Gives another choice for optimization algorithms.
SVM hard-margin duality.

Define the two optimization problems

\[
\min \left\{ \frac{1}{2} \| w \|^2 : w \in \mathbb{R}^d, \forall i : 1 - y_i x_i w \leq 0 \right\} \quad \text{(primal)},
\]

\[
\max \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j : \alpha \in \mathbb{R}^n, \alpha \geq 0 \right\} \quad \text{(dual)}.
\]

If the primal is feasible, then primal optimal value = dual optimal value. Given a primal optimum \( \bar{w} \) and a dual optimum \( \bar{\alpha} \), then

\[
\bar{w} = \sum_{i=1}^{n} \bar{\alpha}_i y_i x_i.
\]
SVM hard-margin duality.

Define the two optimization problems

\[
\begin{align*}
\min & \left\{ \frac{1}{2} \| w \|^2 : w \in \mathbb{R}^d, \forall i \cdot 1 - y_i x_i w \leq 0 \right\} \quad \text{(primal)}, \\
\max & \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j : \alpha \in \mathbb{R}^n, \alpha \geq 0 \right\} \quad \text{(dual)}.
\end{align*}
\]

If the primal is feasible, then primal optimal value = dual optimal value. Given a primal optimum \( \bar{w} \) and a dual optimum \( \bar{\alpha} \), then

\[
\bar{w} = \sum_{i=1}^{n} \bar{\alpha}_i y_i x_i.
\]

▶ The dual variables \( \alpha \) have dimension \( n \), same as examples.
▶ We can write the primal optimum as a linear combination of examples.
▶ The dual objective is a concave quadratic.
▶ We will derive this duality using Lagrange multipliers.
Lagrange multipliers

Move constraints to objective using **Lagrange multipliers**.

**Original problem:**
\[
\min_{w \in \mathbb{R}^d} \quad \frac{1}{2} \|w\|_2^2 \\
\text{s.t.} \quad 1 - y_i x_i^T w \leq 0 \quad \text{for all } i = 1, \ldots, n.
\]

**Lagrangian** \(L(w, \alpha)\):
\[
L(w, \alpha) := \frac{1}{2} \|w\|_2^2 + \sum_{i=1}^n \alpha_i (1 - y_i x_i^T w).
\]

Maximizing over \(\alpha \geq 0\) recovers primal problem: for any \(w \in \mathbb{R}^d\),
\[
P(w) := \sup_{\alpha \geq 0} L(w, \alpha) = \begin{cases} 
\frac{1}{2} \|w\|_2^2 & \text{if } \min_i y_i x_i^T w \geq 1, \\
\infty & \text{otherwise}.
\end{cases}
\]
Lagrange multipliers

Move constraints to objective using Lagrange multipliers.

Original problem:  \[
\min_{\bm{w} \in \mathbb{R}^d} \quad \frac{1}{2} \|\bm{w}\|_2^2
\]
\[
\text{s.t.} \quad 1 - y_i \bm{x}_i^\top \bm{w} \leq 0 \quad \text{for all } i = 1, \ldots, n.
\]

For each constraint \(1 - y_i \bm{x}_i^\top \bm{w} \leq 0\), associate a dual variable (Lagrange multiplier) \(\alpha_i \geq 0\).

Lagrangian \(L(\bm{w}, \alpha)\):

\[
L(\bm{w}, \alpha) := \frac{1}{2} \|\bm{w}\|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i \bm{x}_i^\top \bm{w}).
\]

Maximizing over \(\alpha \geq 0\) recovers primal problem: for any \(\bm{w} \in \mathbb{R}^d\),

\[
P(\bm{w}) := \sup_{\alpha \geq 0} L(\bm{w}, \alpha) = \begin{cases} 
\frac{1}{2} \|\bm{w}\|_2^2 & \text{if } \min_i y_i \bm{x}_i^\top \bm{w} \geq 1, \\
\infty & \text{otherwise}.
\end{cases}
\]

What if we leave \(\alpha\) fixed, and minimize \(\bm{w}\)?
Lagrange multipliers

Move constraints to objective using Lagrange multipliers.

Original problem:

\[
\begin{align*}
\min_{w \in \mathbb{R}^d} & \quad \frac{1}{2} \|w\|_2^2 \\
\text{s.t.} & \quad 1 - y_i x_i^T w \leq 0 \quad \text{for all } i = 1, \ldots, n.
\end{align*}
\]

- For each constraint \(1 - y_i x_i^T w \leq 0\), associate a dual variable (Lagrange multiplier) \(\alpha_i \geq 0\).
- Move constraints to objective by adding \(\sum_{i=1}^{n} \alpha_i (1 - y_i x_i^T w)\) and maximizing over \(\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n\) s.t. \(\alpha \geq 0\) (i.e., \(\alpha_i \geq 0\) for all \(i\)).

Lagrangian \(L(w, \alpha)\):

\[
L(w, \alpha) := \frac{1}{2} \|w\|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i x_i^T w).
\]

Maximizing over \(\alpha \geq 0\) recovers primal problem: for any \(w \in \mathbb{R}^d\),

\[
P(w) := \sup_{\alpha \geq 0} L(w, \alpha) = \begin{cases} 
\frac{1}{2} \|w\|_2^2 & \text{if } \min_i y_i x_i^T w \geq 1, \\
\infty & \text{otherwise}.
\end{cases}
\]

What if we leave \(\alpha\) fixed, and minimize \(w\)?
Dual problem}

Lagrangian

\[ L(w, \alpha) := \frac{1}{2} \|w\|^2_2 + \sum_{i=1}^{n} \alpha_i (1 - y_i x_i^T w). \]

Primal hard-margin SVM

\[ P(w) = \sup_{\alpha \geq 0} L(w, \alpha) = \sup_{\alpha \geq 0} \left[ \frac{1}{2} \|w\|^2_2 + \sum_{i=1}^{n} \alpha_i (1 - y_i x_i^T w) \right]. \]
Dual problem

Lagrangian

\[ L(w, \alpha) := \frac{1}{2} \|w\|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i x_i^T w). \]

Primal hard-margin SVM

\[ P(w) = \sup_{\alpha \geq 0} L(w, \alpha) = \sup_{\alpha \geq 0} \left[ \frac{1}{2} \|w\|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i x_i^T w) \right]. \]

Dual problem \( D(\alpha) = \min_w L(w, \alpha) \): given \( \alpha \geq 0 \), then \( w \mapsto L(w, \alpha) \) is a convex quadratic with minimum \( w = \sum_{i=1}^{n} \alpha_i y_i x_i \), giving

\[ D(\alpha) = \min_{w \in \mathbb{R}^d} L(w, \alpha) = L \left( \sum_{i=1}^{n} \alpha_i y_i x_i, \alpha \right) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \left\| \sum_{i=1}^{n} \alpha_i y_i x_i \right\|_2^2 \]

\[ = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j. \]
Summarizing,

\[ L(w, \alpha) = \frac{1}{2} \|w\|^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i x_i^T w) \]

Lagrangian, \[ P(w) = \max_{\alpha \geq 0} L(w, \alpha) \]
primal problem, \[ D(\alpha) = \min_w L(w, \alpha) \]
dual problem.
Summarizing,

\[ L(w, \alpha) = \frac{1}{2} \|w\|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i x_i^T w) \]  

Lagrangian,

\[ P(w) = \max_{\alpha \geq 0} L(w, \alpha) \]  

primal problem,

\[ D(\alpha) = \min_w L(w, \alpha) \]  

dual problem.

For general Lagrangians, have **weak duality**

\[ P(w) \geq D(\alpha), \]

since \[ P(w) = \max_{\alpha' \geq 0} L(w, \alpha') \geq L(w, \alpha) \geq \min_{w'} L(w', \alpha) = D(\alpha). \]
Summarizing,

\[ L(w, \alpha) = \frac{1}{2} \|w\|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i x_i^T w) \]

Lagrangian,

\[ P(w) = \max_{\alpha \geq 0} L(w, \alpha) \]

primal problem,

\[ D(\alpha) = \min_w L(w, \alpha) \]

dual problem.

▶ For general Lagrangians, have weak duality

\[ P(w) \geq D(\alpha), \]

since \( P(w) = \max_{\alpha' \geq 0} L(w, \alpha') \geq L(w, \alpha) \geq \min_w L(w', \alpha) = D(\alpha). \)

▶ By convexity, have strong duality \( \min_w P(w) = \max_{\alpha \geq 0} D(\alpha), \)

and an optimum \( \bar{\alpha} \) for \( D \) gives an optimum \( \bar{w} \) for \( P \) via

\[ \bar{w} = \sum_{i=1}^{n} \bar{\alpha}_i y_i x_i = \arg \min_w L(w, \bar{\alpha}). \]
Optimal solutions $\tilde{w}$ and $\tilde{\alpha} = (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_n)$ satisfy

$\tilde{w} = \sum_{i=1}^{n} \tilde{\alpha}_i y_i x_i = \sum_{i: \tilde{\alpha}_i > 0} \tilde{\alpha}_i y_i x_i,$

$\tilde{\alpha}_i > 0 \Rightarrow y_i x_i^\top \tilde{w} = 1$ for all $i = 1, \ldots, n$ (complementary slackness).
Optimal solutions $\bar{w}$ and $\bar{\alpha} = (\bar{\alpha}_1, \ldots, \bar{\alpha}_n)$ satisfy

- $\bar{w} = \sum_{i=1}^{n} \bar{\alpha}_i y_i x_i = \sum_{i: \bar{\alpha}_i > 0} \bar{\alpha}_i y_i x_i$,

- $\bar{\alpha}_i > 0 \Rightarrow y_i x_i^T \bar{w} = 1$ for all $i = 1, \ldots, n$ (complementary slackness).

The $y_i x_i$ where $\bar{\alpha}_i > 0$ are called support vectors.

Primal optimum is a linear combination of support vectors.
Optimal solutions $\bar{w}$ and $\bar{\alpha} = (\bar{\alpha}_1, \ldots, \bar{\alpha}_n)$ satisfy

\[
\bar{w} = \sum_{i=1}^{n} \bar{\alpha}_i y_i x_i = \sum_{i: \bar{\alpha}_i > 0} \bar{\alpha}_i y_i x_i,
\]

\[
\bar{\alpha}_i > 0 \implies y_i x_i^T \bar{w} = 1 \text{ for all } i = 1, \ldots, n \quad (\text{complementary slackness}).
\]

The $y_i x_i$ where $\bar{\alpha}_i > 0$ are called support vectors.

- Support vector examples satisfy “margin” constraints with equality.

- Primal optimum is a linear combination of support vectors.
Optimal solutions $\tilde{w}$ and $\tilde{\alpha} = (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_n)$ satisfy

- $\tilde{w} = \sum_{i=1}^{n} \tilde{\alpha}_i y_i x_i = \sum_{i: \tilde{\alpha}_i > 0} \tilde{\alpha}_i y_i x_i$,

- $\tilde{\alpha}_i > 0 \Rightarrow y_i x_i^T \tilde{w} = 1$ for all $i = 1, \ldots, n$ (complementary slackness).

The $y_i x_i$ where $\tilde{\alpha}_i > 0$ are called \textit{support vectors}.

- Support vector examples satisfy “margin” constraints with equality.
- Get same solution if non-support vectors deleted.
- Primal optimum is a linear combination of support vectors.
Proof of complementary slackness

For the optimal (feasible) solutions \( \hat{w} \) and \( \hat{\alpha} \), we have

\[
P(\hat{w}) = D(\hat{\alpha}) = \min_{w \in \mathbb{R}^d} L(w, \hat{\alpha}) \quad \text{(by strong duality)}
\]
Proof of complementary slackness

For the optimal (feasible) solutions \( \hat{w} \) and \( \hat{\alpha} \), we have

\[
P(\hat{w}) = D(\hat{\alpha}) = \min_{\mathbf{w} \in \mathbb{R}^d} L(\mathbf{w}, \hat{\alpha}) \quad \text{(by strong duality)}
\]

\[
\leq L(\hat{w}, \hat{\alpha})
\]
Proof of complementary slackness

For the optimal (feasible) solutions \( \hat{w} \) and \( \hat{\alpha} \), we have

\[
P(\hat{w}) = D(\hat{\alpha}) = \min_{w \in \mathbb{R}^d} L(w, \hat{\alpha}) \quad \text{(by strong duality)}
\]

\[
\leq L(\hat{w}, \hat{\alpha})
\]

\[
= \frac{1}{2} \|\hat{w}\|_2^2 + \sum_{i=1}^{n} \hat{\alpha}_i (1 - y_i x_i^T \hat{w})
\]
Proof of complementary slackness

For the optimal (feasible) solutions \( \hat{w} \) and \( \hat{\alpha} \), we have

\[
P(\hat{w}) = D(\hat{\alpha}) = \min_{w \in \mathbb{R}^d} L(w, \hat{\alpha}) \quad \text{(by strong duality)}
\]

\[
\leq L(\hat{w}, \hat{\alpha})
\]

\[
= \frac{1}{2} \|\hat{w}\|_2^2 + \sum_{i=1}^{n} \hat{\alpha}_i (1 - y_i \hat{x}_i^T \hat{w})
\]

\[
\leq \frac{1}{2} \|\hat{w}\|_2^2 \quad \text{(constraints are satisfied)}
\]

\[
= P(\hat{w}).
\]
Proof of complementary slackness

For the optimal (feasible) solutions \( \hat{w} \) and \( \hat{\alpha} \), we have

\[
P(\hat{w}) = D(\hat{\alpha}) = \min_{\mathbf{w} \in \mathbb{R}^d} L(\mathbf{w}, \hat{\alpha}) \quad \text{(by strong duality)}
\]

\[
\leq L(\hat{w}, \hat{\alpha})
\]

\[
= \frac{1}{2} \|\hat{w}\|_2^2 + \sum_{i=1}^{n} \hat{\alpha}_i (1 - y_i \mathbf{x}_i^T \hat{w})
\]

\[
\leq \frac{1}{2} \|\hat{w}\|_2^2 \quad \text{(constraints are satisfied)}
\]

\[
= P(\hat{w}).
\]

Therefore, every term in sum \( \sum_{i=1}^{n} \hat{\alpha}_i (1 - y_i \mathbf{x}_i^T \hat{w}) \) must be zero:

\[
\hat{\alpha}_i (1 - y_i \mathbf{x}_i^T \hat{w}) = 0 \quad \text{for all } i = 1, \ldots, n.
\]
Proof of complementary slackness

For the optimal (feasible) solutions $\hat{w}$ and $\hat{\alpha}$, we have

$$\begin{align*}
P(\hat{w}) &= D(\hat{\alpha}) = \min_{w \in \mathbb{R}^d} L(w, \hat{\alpha}) \quad \text{(by strong duality)} \\
&\leq L(\hat{w}, \hat{\alpha}) \\
&= \frac{1}{2} \Vert \hat{w} \Vert_2^2 + \sum_{i=1}^{n} \hat{\alpha}_i (1 - y_i \mathbf{x}_i^T \hat{w}) \\
&\leq \frac{1}{2} \Vert \hat{w} \Vert_2^2 \quad \text{(constraints are satisfied)} \\
&= P(\hat{w}).
\end{align*}$$

Therefore, every term in sum $\sum_{i=1}^{n} \hat{\alpha}_i (1 - y_i \mathbf{x}_i^T \hat{w})$ must be zero:

$$\hat{\alpha}_i (1 - y_i \mathbf{x}_i^T \hat{w}) = 0 \quad \text{for all } i = 1, \ldots, n.$$ 

If $\alpha_i > 0$, then must have $1 - y_i \mathbf{x}_i^T \hat{w} = 0$. 

Lagrangian

\[ L(w, \alpha) = \frac{1}{2} \|w\|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i x_i^T w). \]

Primal maximum margin problem was

\[ P(w) = \sup_{\alpha \geq 0} L(w, \alpha) = \sup_{\alpha \geq 0} \left[ \frac{1}{2} \|w\|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i x_i^T w) \right]. \]

Dual problem

\[ D(\alpha) = \min_{w \in \mathbb{R}^d} L(w, \alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \left\| \sum_{i=1}^{n} \alpha_i y_i x_i \right\|_2^2. \]

Given dual optimum \( \bar{\alpha} \),

- Corresponding primal optimum \( \bar{w} = \sum_{i=1}^{n} \alpha_i y_i x_i \);
- Strong duality \( P(\bar{w}) = D(\bar{\alpha}) \);
- \( \bar{\alpha}_i > 0 \) implies \( y_i x_i^T \bar{w} = 1 \), and these \( y_i x_i \) are support vectors.
Similarly,

\[
L(w, \xi, \alpha) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i (1 - \xi_i - y_i x_i^T w)
\]  

(Lagrangian),

\[
P(w, \xi) = \sup_{\alpha \geq 0} L(w, \xi, \alpha)
\]  

(Primal),

\[
D(\alpha) = \min_{w \in \mathbb{R}^d, \xi \in \mathbb{R}^n_\geq 0} L(w, \xi, \alpha)
\]  

(Dual),

\[
= \max_{\alpha \in \mathbb{R}^n, 0 \leq \alpha_i \leq C} \left[ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \left\| \sum_{i=1}^{n} \alpha_i y_i x_i \right\|^2 \right].
\]

Remarks.

▶ Dual solution \(\bar{\alpha}\) still gives primal solution \(\bar{w}\) = \(\sum_{i=1}^{n} \bar{\alpha}_i y_i x_i\).

▶ Can take \(C \to \infty\) to recover hard-margin case.

▶ Dual is still a constrained convex quadratic (used in many solvers).

▶ Some presentations include bias in primal (\(x_i^T w + b\)); this introduces a constraint \(\sum_{i=1}^{n} y_i \alpha_i = 0\) in dual.
Similarly,

\[ L(w, \xi, \alpha) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i (1 - \xi_i - y_i x_i^T w) \]  

(Lagrangian),

\[ P(w, \xi) = \sup_{\alpha \geq 0} L(w, \xi, \alpha) \]  

(Primal),

\[ D(\alpha) = \min_{w \in \mathbb{R}^d, \xi \in \mathbb{R}^n \geq 0} L(w, \xi, \alpha) \]  

(Dual),

\[ = \begin{cases} 
\frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i & \forall i \cdot 1 - \xi_i - y_i x_i^T w \leq 0, \\
\infty & \text{otherwise}, 
\end{cases} \]
Convex dual in non-separable case

Similarly,

\[ L(w, \xi, \alpha) = \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i (1 - \xi_i - y_i x_i^T w) \]  

(Lagrangian),

\[ P(w, \xi) = \sup_{\alpha \geq 0} L(w, \xi, \alpha) \]  

(Primal),

\[ = \begin{cases} 
\frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^{n} \xi_i & \forall i \cdot 1 - \xi_i - y_i x_i^T w \leq 0, \\
\infty & \text{otherwise,}
\end{cases} \]

\[ D(\alpha) = \min_{w \in \mathbb{R}^d, \xi \in \mathbb{R}^n_{\geq 0}} L(w, \xi, \alpha) \]  

(Dual),

\[ = \max_{\alpha \in \mathbb{R}^n} \left[ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \left\| \sum_{i=1}^{n} \alpha_i y_i x_i \right\|_2^2 \right]. \]

Remarks.

- Dual solution \( \bar{\alpha} \) still gives primal solution \( \bar{w} = \sum_{i=1}^{n} \bar{\alpha}_i y_i x_i \).
- Can take \( C \to \infty \) to recover hard-margin case.
- Dual is still a constrained convex quadratic (used in many solvers).
- Some presentations include bias in primal \( x_i^T w + b \); this introduces a constraint \( \sum_{i=1}^{n} y_i \alpha_i = 0 \) in dual.
Nonlinear SVM: feature mapping annoying in the primal?

SVM hard-margin primal, with a feature mapping $\phi$: $\mathbb{R}^d \rightarrow \mathbb{R}^p$:

$$\min \left\{ \frac{1}{2} \| w \|_2^2 : w \in \mathbb{R}^p, \forall i \phi(x_i)^T w \geq 1 \right\}.$$ 

Now the search space has $p$ dimensions; potentially $p \gg d$.

Can we do better?
Nonlinear SVM: feature mapping annoying in the primal?

SVM hard-margin primal, with a feature mapping $\phi : \mathbb{R}^d \to \mathbb{R}^p$:

$$\min \left\{ \frac{1}{2} \| w \|^2 : w \in \mathbb{R}^p, \forall i \cdot \phi(x_i)^T w \geq 1 \right\}.$$

Now the search space has $p$ dimensions; potentially $p \gg d$.

Can we do better?
Feature mapping in the dual

Given dual optimum $\bar{\alpha}$, since $\bar{w} = \sum_{i=1}^{n} \bar{\alpha}_i y_i \phi(x_i)$, we can predict on future $x$ with $x \mapsto \phi(x) \transp \bar{w} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \phi(x) \transp \phi(x_i)$.

▶ Dual form never needs $\phi(x) \in \mathbb{R}^p$, only $\phi(x) \transp \phi(x_i) \in \mathbb{R}$.

▶ Kernel trick: replace every $\phi(x) \transp \phi(x')$ with kernel evaluation $k(\cdot, \cdot)$.

Sometimes $k(\cdot, \cdot)$ is much cheaper than $\phi(x) \transp \phi(x')$.

▶ This idea started with SVM, but appears in many other places.

▶ Downside: implementations usuall store Gram matrix $G \in \mathbb{R}^{n \times n}$ where $G_{ij} := k(x_i, x_j)$. 

24 / 37
Feature mapping in the dual

SVM hard-margin dual, with a feature mapping \( \phi : \mathbb{R}^d \rightarrow \mathbb{R}^p \):

\[
\max_{\alpha_1, \alpha_2, \ldots, \alpha_n \geq 0} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \phi(x_i)^T \phi(x_j).
\]

Given dual optimum \( \bar{\alpha} \), since \( \bar{w} = \sum_{i=1}^{n} \bar{\alpha}_i y_i \phi(x_i) \), we can predict on future \( x \) with

\[
\bar{w} = \sum_{i=1}^{n} \bar{\alpha}_i y_i \phi(x_i)^T \phi(x_i).
\]
Feature mapping in the dual

SVM hard-margin dual, with a feature mapping \( \phi : \mathbb{R}^d \rightarrow \mathbb{R}^p \):

\[
\max_{\alpha_1, \alpha_2, \ldots, \alpha_n \geq 0} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \phi(x_i)^T \phi(x_j).
\]

Given dual optimum \( \bar{\alpha} \), since \( \bar{w} = \sum_{i=1}^{n} \bar{\alpha}_i y_i \phi(x_i) \), we can predict on future \( x \) with

\[
x \mapsto \phi(x)^T \bar{w} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \phi(x)^T \phi(x_i).
\]

- **Dual form never needs** \( \phi(x) \in \mathbb{R}^p \), only \( \phi(x)^T \phi(x_i) \in \mathbb{R} \).
- **Kernel trick:** replace every \( \phi(x)^T \phi(x') \) with kernel evaluation \( k(x, x') \). Sometimes \( k(\cdot, \cdot) \) is much cheaper than \( \phi(x)^T \phi(x') \).
- This idea started with SVM, but appears in many other places.
- **Downside:** implementations usuall store Gram matrix \( G \in \mathbb{R}^{n \times n} \) where \( G_{ij} := k(x_i, x_j) \).
Affine features: $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^{1+d}$, where

$$\phi(x) = (1, x_1, \ldots, x_d).$$

Kernel form:

$$\phi(x)^T \phi(x') = 1 + x^T x'.$$
Kernel example: quadratic features

Consider re-normalized quadratic features
\( \phi : \mathbb{R}^d \rightarrow \mathbb{R}^{1+2d+\binom{d}{2}} \), where

\[
\phi(\mathbf{x}) = \left(1, \sqrt{2}x_1, \ldots, \sqrt{2}x_d, x_1^2, \ldots, x_d^2, \sqrt{2}x_1x_2, \ldots, \sqrt{2}x_1x_d, \ldots, \sqrt{2}x_{d-1}x_d \right).
\]

Just writing this down takes time \( \mathcal{O}(d^2) \).
Meanwhile,

\[
\phi(\mathbf{x})^\top \phi(\mathbf{x}') = (1 + \mathbf{x}^\top \mathbf{x}')^2,
\]

time \( \mathcal{O}(d) \).
Consider re-normalized quadratic features
\( \phi: \mathbb{R}^d \to \mathbb{R}^{1+2d+\binom{d}{2}} \), where

\[
\phi(x) = \left( 1, \sqrt{2}x_1, \ldots, \sqrt{2}x_d, x_1^2, \ldots, x_d^2, \sqrt{2}x_1x_2, \ldots, \sqrt{2}x_1x_d, \ldots, \sqrt{2}x_{d-1}x_d \right).
\]

Just writing this down takes time \( O(d^2) \).
Meanwhile,

\[
\phi(x)^T \phi(x') = (1 + x^T x')^2,
\]

time \( O(d) \).

**Tweaks:**

- What if we change exponent “2”?
- What if we replace additive “1” with 0?
Consider \( \phi: \mathbb{R}^d \to \mathbb{R}^{2^d} \), where

\[
\phi(x) = \left( \prod_{i \in S} x_i \right)_{S \subseteq \{1,2,\ldots,d\}}
\]

Time \( O(2^d) \) just to write down. Kernel evaluation takes time \( O(d) \):

\[
\phi(x)^T \phi(x') = \prod_{i=1}^{d} (1 + x_i x'_i).
\]
For any $\sigma > 0$, there is an infinite feature expansion $\phi: \mathbb{R}^d \rightarrow \mathbb{R}^\infty$ such that

$$\phi(x)^T \phi(x') = \exp \left( -\frac{\|x - x'\|^2}{2\sigma^2} \right),$$

which can be computed in $O(d)$ time.

This is called a Gaussian kernel or RBF kernel. It has some similarities to nearest neighbor methods (later lecture).

$\phi$ maps to an infinite-dimensional space, but there’s no reason to know that.
Defining kernels without $\phi$

A (positive definite) **kernel function** $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is a symmetric function so that for any $n$ and any data examples $(\mathbf{x}_i)_{i=1}^n$, the corresponding Gram matrix $G \in \mathbb{R}^{n \times n}$ with $G_{ij} := k(\mathbf{x}_i, \mathbf{x}_j)$ is positive semi-definite.
A (positive definite) kernel function \( k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R} \) is a symmetric function so that for any \( n \) and any data examples \((x_i)_{i=1}^n\), the corresponding Gram matrix \( G \in \mathbb{R}^{n \times n} \) with \( G_{ij} := k(x_i, x_j) \) is positive semi-definite.

- There is a ton of theory about this formalism; e.g., keywords RKHS, representer theorem, Mercer’s theorem.
- Given any such \( k \), there always exists a corresponding \( \phi \).
- This definition ensures the SVM dual is still concave.
Source data.

Quadratic SVM.

RBF SVM ($\sigma = 1$).

RBF SVM ($\sigma = 0.1$).
Summary for SVM

- Hard-margin SVM.
- Soft-margin SVM.
- SVM duality.
- Nonlinear SVM: kernels
(Appendix.)
Let’s derive the final dual form:

\[
L(w, \xi, \alpha) = \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i (1 - \xi_i - y_i x_i^T w)
\]

\[
D(\alpha) = \min_{w \in \mathbb{R}^d, \xi \in \mathbb{R}^n \geq 0} L(w, \xi, \alpha) = \max_{\alpha \in \mathbb{R}^n} \left[ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \left\| \sum_{i=1}^{n} \alpha_i y_i x_i \right\|^2 \right].
\]

Given \( \alpha \) and \( \xi \), the minimizing \( w \) is still \( w = \sum_{i=1}^{n} \alpha_i y_i x_i \); plugging in,

\[
D(\alpha) = \min_{\xi \in \mathbb{R}^n \geq 0} L \left( \sum_{i=1}^{n} \alpha_i y_i x_i, \xi, \alpha \right) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \left\| \sum_{i=1}^{n} \alpha_i y_i x_i \right\|^2 + \sum_{i=1}^{n} \xi_i (C - \alpha_i).
\]

The goal is to maximize \( D \); if \( \alpha_i > C \), then \( \xi_i \uparrow \infty \) gives \( D(\alpha) = -\infty \). Otherwise, minimized at \( \xi_i = 0 \). Therefore the dual problem is

\[
\max_{\alpha \in \mathbb{R}^n} \left[ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \left\| \sum_{i=1}^{n} \alpha_i y_i x_i \right\|^2 \right]
\]

subject to \( 0 \leq \alpha_i \leq C \).
First consider $d = 1$, meaning $\phi : \mathbb{R} \rightarrow \mathbb{R}^\infty$.

What $\phi$ has $\phi(x)\phi(y) = e^{-(x-y)^2/(2\sigma^2)}$?
First consider $d = 1$, meaning $\phi: \mathbb{R} \to \mathbb{R}^\infty$.

What $\phi$ has $\phi(x)\phi(y) = e^{-(x-y)^2/(2\sigma^2)}$?

Reverse engineer using Taylor expansion:

$$e^{-(x-y)^2/(2\sigma^2)} = e^{-x^2/(2\sigma^2)} \cdot e^{-y^2/(2\sigma^2)} \cdot e^{xy/\sigma^2}$$
Gaussian kernel feature expansion

First consider $d = 1$, meaning $\phi: \mathbb{R} \to \mathbb{R}^\infty$.

What $\phi$ has $\phi(x)\phi(y) = e^{-(x-y)^2/(2\sigma^2)}$?

Reverse engineer using Taylor expansion:

$$e^{-(x-y)^2/(2\sigma^2)} = e^{-x^2/(2\sigma^2)} \cdot e^{-y^2/(2\sigma^2)} \cdot e^{xy/\sigma^2}$$

$$= e^{-x^2/(2\sigma^2)} \cdot e^{-y^2/(2\sigma^2)} \cdot \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{xy}{\sigma^2}\right)^k$$
Gaussian kernel feature expansion

First consider $d = 1$, meaning $\phi: \mathbb{R} \rightarrow \mathbb{R}^\infty$.

What $\phi$ has $\phi(x)\phi(y) = e^{-(x-y)^2/(2\sigma^2)}$?

Reverse engineer using Taylor expansion:

\[
e^{-(x-y)^2/(2\sigma^2)} = e^{-x^2/(2\sigma^2)} \cdot e^{-y^2/(2\sigma^2)} \cdot e^{xy/\sigma^2}
\]

\[
= e^{-x^2/(2\sigma^2)} \cdot e^{-y^2/(2\sigma^2)} \cdot \sum_{k=0}^{\infty} \frac{1}{k!} \left( \frac{xy}{\sigma^2} \right)^k
\]

So let

\[
\phi(x) := e^{-x^2/(2\sigma^2)} \left( 1, \frac{x}{\sigma}, \frac{1}{\sqrt{2!}} \left( \frac{x}{\sigma} \right)^2, \frac{1}{\sqrt{3!}} \left( \frac{x}{\sigma} \right)^3, \cdots \right).
\]
Gaussian kernel feature expansion

First consider \( d = 1 \), meaning \( \phi : \mathbb{R} \to \mathbb{R}^\infty \).

What \( \phi \) has \( \phi(x)\phi(y) = e^{-(x-y)^2/(2\sigma^2)} \)?

Reverse engineer using Taylor expansion:

\[
e^{-\frac{(x-y)^2}{2\sigma^2}} = e^{-\frac{x^2}{2\sigma^2}} \cdot e^{-\frac{y^2}{2\sigma^2}} \cdot e^{\frac{xy}{\sigma^2}}
= e^{-\frac{x^2}{2\sigma^2}} \cdot e^{-\frac{y^2}{2\sigma^2}} \cdot \sum_{k=0}^{\infty} \frac{1}{k!} \left( \frac{xy}{\sigma^2} \right)^k
\]

So let

\[
\phi(x) := e^{-\frac{x^2}{2\sigma^2}} \left( 1, \frac{x}{\sigma}, \frac{1}{\sqrt{2!}} \left( \frac{x}{\sigma} \right)^2, \frac{1}{\sqrt{3!}} \left( \frac{x}{\sigma} \right)^3, \ldots \right).
\]

How to handle \( d > 1 \)?
Gaussian kernel feature expansion

First consider \( d = 1 \), meaning \( \phi : \mathbb{R} \to \mathbb{R}^\infty \).

What \( \phi \) has \( \phi(x)\phi(y) = e^{-(x-y)^2/(2\sigma^2)} \)?

Reverse engineer using Taylor expansion:

\[
e^{-\frac{(x-y)^2}{2\sigma^2}} = e^{-\frac{x^2}{2\sigma^2}} \cdot e^{-\frac{y^2}{2\sigma^2}} \cdot e^{\frac{xy}{\sigma^2}}
\]

\[
= e^{-\frac{x^2}{2\sigma^2}} \cdot e^{-\frac{y^2}{2\sigma^2}} \cdot \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{xy}{\sigma^2}\right)^k
\]

So let

\[
\phi(x) := e^{-\frac{x^2}{2\sigma^2}} \left(1, \frac{x}{\sigma}, \frac{1}{\sqrt{2}} \left(\frac{x}{\sigma}\right)^2, \frac{1}{\sqrt{3}} \left(\frac{x}{\sigma}\right)^3, \ldots\right).
\]

How to handle \( d > 1 \)?

\[
e^{-\|x-y\|^2/(2\sigma^2)} = e^{-\|x\|^2/(2\sigma^2)} \cdot e^{-\|y\|^2/(2\sigma^2)} \cdot e^{\frac{x^T y}{\sigma^2}}
\]

\[
= e^{-\|x\|^2/(2\sigma^2)} \cdot e^{-\|y\|^2/(2\sigma^2)} \cdot \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{x^T y}{\sigma^2}\right)^k
\]
Other kernels

Suppose $k_1$ and $k_2$ are positive definite kernel functions.

Another approach: random features $k(x, x') = \mathbb{E}_w F(w, x)^\top F(w, x')$ for some $F$; we will revisit this with deep networks and the neural tangent kernel (NTK).
Suppose $k_1$ and $k_2$ are positive definite kernel functions.

1. $k(x, y) := k_1(x, y) + k_2(x, y)$ define a positive definite kernel?

Another approach: random features $k(x, x') = \mathbb{E}_w F(w, x')^\top F(w, x')$ for some $F$; we will revisit this with deep networks and the neural tangent kernel (NTK).
Other kernels

Suppose $k_1$ and $k_2$ are positive definite kernel functions.

1. $k(x, y) := k_1(x, y) + k_2(x, y)$ define a positive definite kernel?

2. $k(x, y) := c \cdot k_1(x, y)$ (for $c \geq 0$) define a positive definite kernel?

Another approach: random features $k(x, x') = \mathbb{E}_w F(w, x')^\top F(w, x')$ for some $F$; we will revisit this with deep networks and the neural tangent kernel (NTK).
Suppose $k_1$ and $k_2$ are positive definite kernel functions.

1. $k(x, y) := k_1(x, y) + k_2(x, y)$ define a positive definite kernel?
2. $k(x, y) := c \cdot k_1(x, y)$ (for $c \geq 0$) define a positive definite kernel?
3. $k(x, y) := k_1(x, y) \cdot k_2(x, y)$ define a positive definite kernel?

Another approach: random features $k(x, x') = \mathbb{E}_w F(w, x')^\top F(w, x')$ for some $F$; we will revisit this with deep networks and the neural tangent kernel (NTK).
Kernel ridge regression:

\[
\min_w \frac{1}{2n} \|Xw - y\|^2 + \frac{\lambda}{2} \|w\|^2.
\]

Solution:

\[
\hat{w} = (X^TX + \lambda nI)^{-1} X^T y.
\]

Linear algebra fact:

\[
X^T (XX^T + \lambda nI)^{-1} y.
\]

Therefore predict with

\[
x \mapsto x^T \hat{w} = (Xx)^T (XX^T + \lambda nI)^{-1} y = \sum_{i=1}^n (x_i^T x)^T \left[ (XX^T + \lambda nI)^{-1} y \right]_i.
\]

Kernel approach:

- Compute \( \alpha := (G + \lambda nI)^{-1} \), where \( G \in \mathbb{R}^n \) is the Gram matrix:
  \[
  G_{ij} = k(x_i, x_j).
  \]
- Predict with \( x \mapsto \sum_{i=1}^n \alpha_i x_i^T x \).
Multiclass SVM

There are a few ways; one is to use one-against-all as in lecture.

Many researchers have proposed various multiclass SVM methods, but some of them can be shown to fail in trivial cases.