CSCI 567: Machine Learning

Vatsal Sharan

Fall 2022

Lecture 5, Sep 22

USC University of Southern California

Administrivia

- HW2 due in about a week.
- Quiz 1 in 2 weeks.

Recap

Regularized least squares

We looked at regularized least squares with non-linear basis:

$$\begin{aligned} \boldsymbol{w}^* &= \operatorname*{argmin}_{\boldsymbol{w}} F(\boldsymbol{w}) \\ &= \operatorname*{argmin}_{\boldsymbol{w}} \left(\|\boldsymbol{\Phi}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \lambda \|\boldsymbol{w}\|_2^2 \right) \\ &= \left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \lambda \boldsymbol{I} \right)^{-1} \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{y} \end{aligned}$$

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi(\boldsymbol{x}_1)^{\mathrm{T}} \\ \phi(\boldsymbol{x}_2)^{\mathrm{T}} \\ \vdots \\ \phi(\boldsymbol{x}_n)^{\mathrm{T}} \end{pmatrix}, \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

This solution operates in the space $\mathbb{R}^M$ and M could be huge (and even infinite).

Regularized least squares solution: Another look

We realized that we can write,

$$\boldsymbol{w}^* = \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\alpha} = \sum_{i=1}^{n} \alpha_i \boldsymbol{\phi}(\boldsymbol{x}_i)$$

Thus the least square solution is **a linear combination of features of the datapoints**!

We calculated what $\boldsymbol{\alpha}$ should be,

$$\boldsymbol{\alpha} = (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1}\boldsymbol{y}$$

where $\boldsymbol{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} \in \mathbb{R}^{n \times n}$ is the **kernel matrix**.

Kernel trick

The prediction of $\boldsymbol{w}^*$ on a new example $\boldsymbol{x}$ is

$$\boldsymbol{w}^{*\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i \boldsymbol{\phi}(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x})$$

Therefore, *only inner products in the new feature space matter!*

Kernel methods are exactly about computing inner products *without explicitly computing ϕ*. The exact form of ϕ is inessential; *all we need to do is know the inner products $\boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{\phi}(\boldsymbol{x}')$.*

The kernel trick: Example 1

Consider the following polynomial basis $\boldsymbol{\phi} : \mathbb{R}^2 \to \mathbb{R}^3$:

$$\boldsymbol{\phi}(\boldsymbol{x}) = \begin{pmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{pmatrix}$$

What is the inner product between $\boldsymbol{\phi}(\boldsymbol{x})$ and $\boldsymbol{\phi}(\boldsymbol{x}')$?

$$\begin{aligned} \boldsymbol{\phi}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}') &= {x_1}^2{x'_1}^2 + 2x_1x_2x'_1x'_2 + {x_2}^2{x'_2}^2 \\ &= (x_1x'_1 + x_2x'_2)^2 = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}')^2 \end{aligned}$$

Therefore, *the inner product in the new space is simply a function of the inner product in the original space*.

Kernel functions

Definition: a function $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is called a *kernel function* if there exists a function $\boldsymbol{\phi} : \mathbb{R}^d \rightarrow \mathbb{R}^M$ so that for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$,

$$k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x})^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}')$$

Popular kernels:

1. Polynomial kernel

$$k(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x}^{\mathrm{T}} \boldsymbol{x}' + c)^M$$

for $c \geq 0$ and M is a positive integer.

2. Gaussian kernel or Radial basis function (RBF) kernel

$$k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2}{2\sigma^2}\right) \quad \text{for some } \sigma > 0.$$

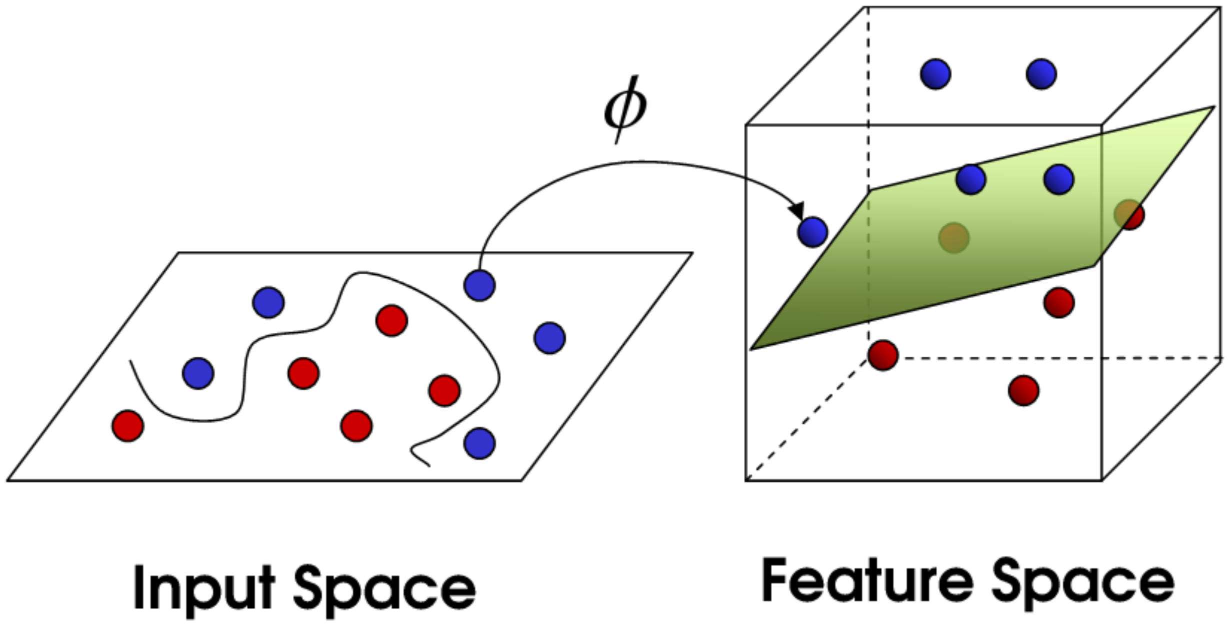
Prediction with kernels

As long as $\boldsymbol{w}^* = \sum_{i=1}^{n} \alpha_i \boldsymbol{\phi}(\boldsymbol{x}_i)$, prediction on a new example $\boldsymbol{x}$ becomes

$$\boldsymbol{w}^{*\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i \boldsymbol{\phi}(\boldsymbol{x}_i)^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x}).$$

This is known as a **non-parametric method**. Informally speaking, this means that there is no fixed set of parameters that the model is trying to learn (remember $\boldsymbol{w}^*$ could be infinite). Nearest-neighbors is another non-parametric method we have seen.

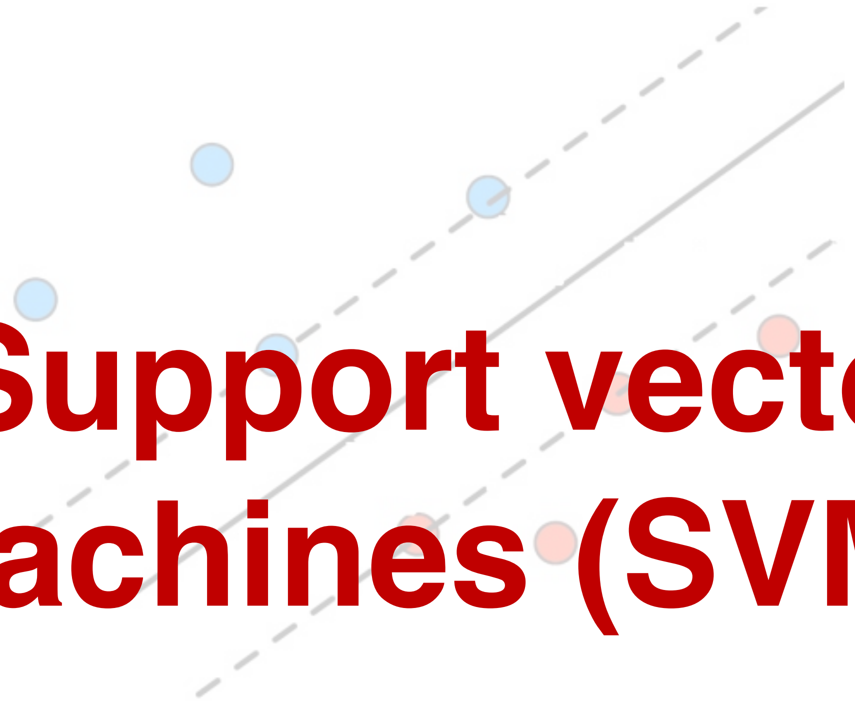
Classification with kernels



Similar ideas extend to the classification case, and we can predict using $\mathrm{sign}(\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}))$.
Data may become linearly separable in the feature space!

We'll see this today.

Support vector machines (SVMs)



1.1 Why study SVM?

- One of the most commonly used classification algorithms
- Allows us to explore the concept of *margins* in classification
- Works well with the kernel trick
- Strong theoretical guarantees

We focus on **binary classification** here.

The *function class for SVMs is a linear function on a feature map ϕ applied to the datapoints*: $\mathrm{sign}(\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}) + b)$. Note, the bias term b is taken separately for SVMs, you'll see why.

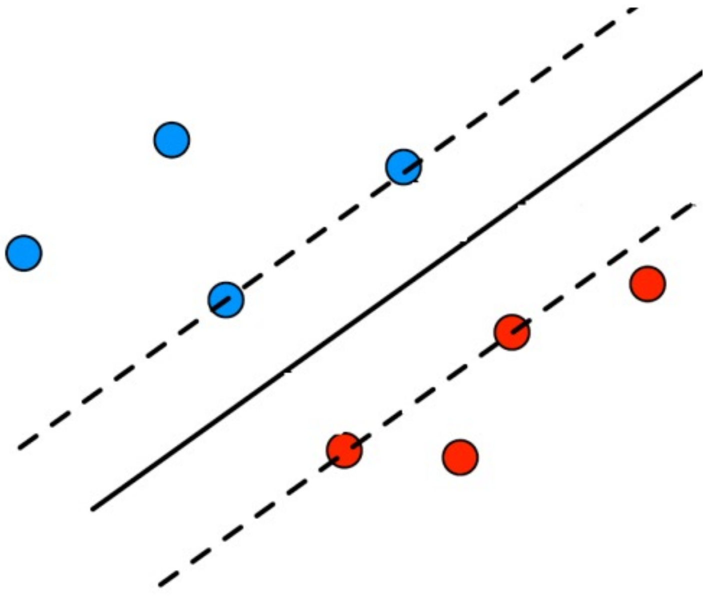
1.2 Margins: separable case, geometric intuition

When data is **linearly separable**, there are infinitely many hyperplanes with zero training error:

Which one should we choose?

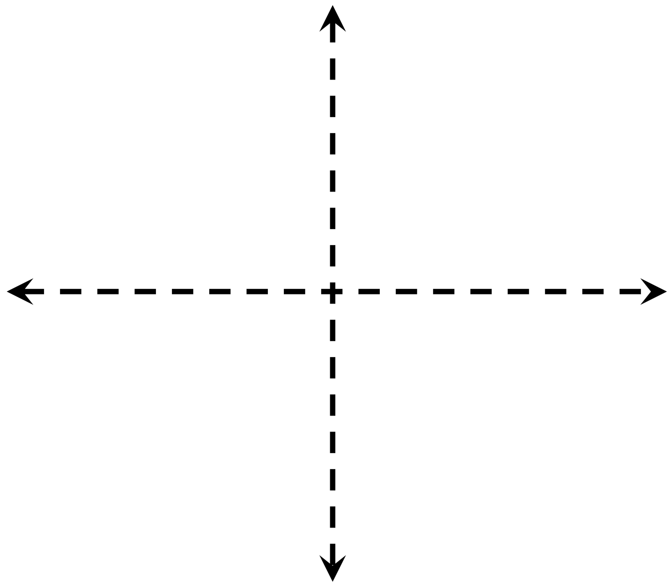
1.2 Margins: separable case, geometric intuition

The further away the separating hyperplane is from the datapoints, the better.



1.2 Formalizing geometric intuition: Distance to hyperplane

What is the **distance** from a point $\boldsymbol{x}$ to a hyperplane $\{\boldsymbol{x} : \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b = 0\}$?



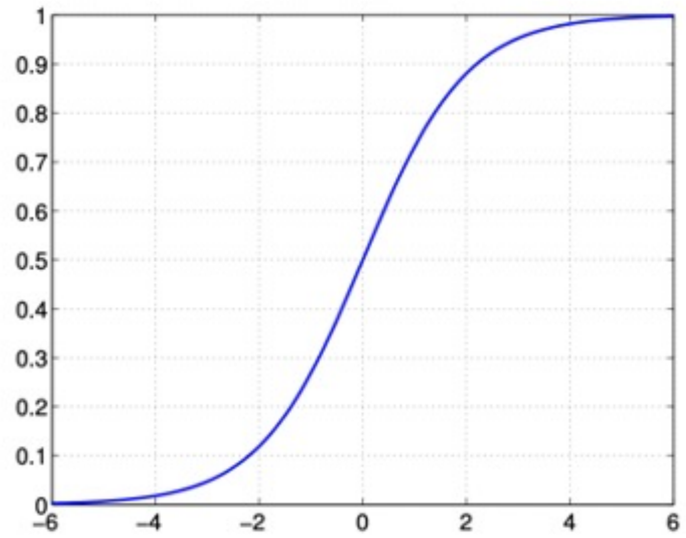
Assume the **projection** is $\boldsymbol{x}' = \boldsymbol{x} - \beta\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}$, then

$$0 = \boldsymbol{w}^{\mathrm{T}}\left(\boldsymbol{x} - \beta\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}\right) + b = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} - \beta\|\boldsymbol{w}\| + b \implies \beta = \frac{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b}{\|\boldsymbol{w}\|_2}.$$

Therefore the distance is $\|\boldsymbol{x} - \boldsymbol{x}'\|_2 = |\beta| = \frac{|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b|}{\|\boldsymbol{w}\|_2}$.

For a hyperplane that correctly classifies $(\boldsymbol{x}, y)$, the distance becomes $\frac{y(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b)}{\|\boldsymbol{w}\|_2}$.

1.2 Margins: functional motivation

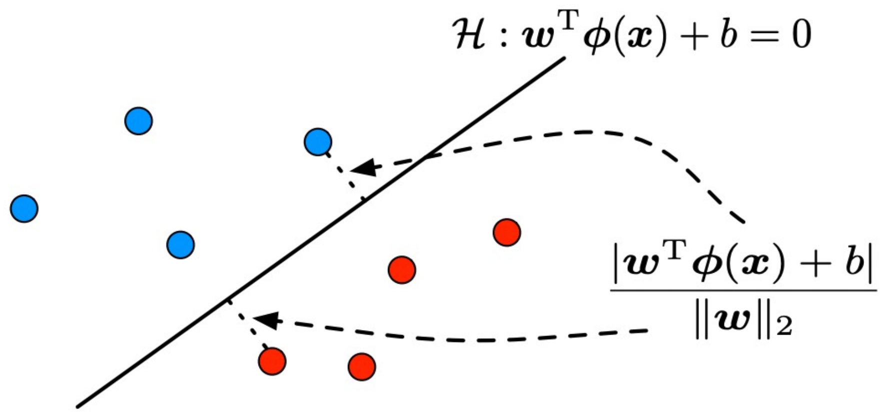


$$\Pr[y \mid \boldsymbol{x}; \boldsymbol{w}] = \sigma(y(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b)) = \frac{1}{1+\exp(-y(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b))}$$

1.3 Maximizing margin

Margin: the *smallest* distance from all training points to the hyperplane

$$\text{MARGIN OF } (\boldsymbol{w}, b) = \min_i \frac{y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_i) + b)}{\|\boldsymbol{w}\|_2}$$



The intuition "**the further away the better**" translates to solving

$$\max_{\boldsymbol{w},b} \; \min_i \frac{y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_i) + b)}{\|\boldsymbol{w}\|_2} = \max_{\boldsymbol{w},b} \frac{1}{\|\boldsymbol{w}\|_2} \min_i y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_i) + b)$$

1.3 Maximizing margin, rescaling

Note: rescaling $(\boldsymbol{w}, b)$ by multiplying both by some scalar does not change the hyperplane.

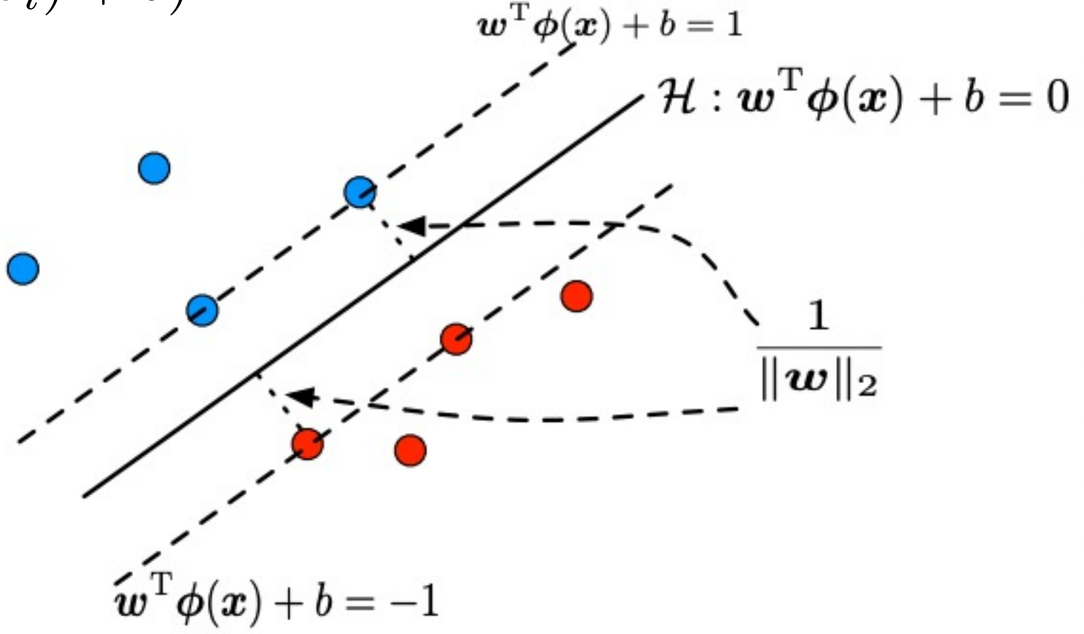
We can thus always scale $(\boldsymbol{w}, b)$ s.t. $\min_i y_i(\boldsymbol{w}^\mathrm{T}\boldsymbol{\phi}(\boldsymbol{x}_i) + b) = 1$

The margin then becomes

MARGIN OF $(\boldsymbol{w}, b)$

$$= \frac{1}{\|\boldsymbol{w}\|_2} \min_i y_i(\boldsymbol{w}^\mathrm{T}\boldsymbol{\phi}(\boldsymbol{x}_i) + b)$$

$$= \frac{1}{\|\boldsymbol{w}\|_2}$$



1.4 SVM for separable data: "Primal" formulation

For a separable training set, we aim to solve

$$\max_{\boldsymbol{w},b} \frac{1}{\|\boldsymbol{w}\|_2} \quad \text{s.t.} \quad \min_i y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_i) + b) = 1$$

This is equivalent to

$$\begin{aligned} \min_{\boldsymbol{w},b} \quad & \frac{1}{2}\|\boldsymbol{w}\|_2^2 \\ \text{s.t.} \quad & y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_i) + b) \geq 1, \ \ \forall\, i \in [n] \end{aligned}$$

SVM is thus also called *max-margin* classifier. The constraints above are called *hard-margin* constraints.

1.5 General non-separable case

If data is not linearly separable, the previous constraint

$$y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_i) + b) \geq 1, \ \ \forall\, i \in [n]$$

is obviously *not feasible*. What is the right thing to do?

1.5 General non-separable case

If data is not linearly separable, the previous constraint $y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_i)+b) \geq 1, \ \forall\, i \in [n]$ is not feasible. And more generally, forcing classifier to always classify all datapoints correctly may not be the best idea.

To deal with this issue, we relax the constraints to ℓ_1 **norm soft-margin** constraints:

$$
\begin{aligned}
& y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_i) + b) \geq 1 - \xi_i, \quad \forall\, i \in [n] \\
\iff \quad & 1 - y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_i) + b) \leq \xi_i, \quad \forall\, i \in [n]
\end{aligned}
$$

where we introduce **slack variables** $\xi_i \geq 0$.

Recall the hinge loss: $\ell_{\text{hinge}}(z) = \max\{0, 1 - z\}$. In our case, $z = y(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}) + b)$.

Aside: Why ℓ_1 penalization?

Aside: Why ℓ_1 penalization?

Aside: Why ℓ_1 penalization?

1.5 Back to SVM: General non-separable case

If data is not linearly separable, the constraint $y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_i) + b) \geq 1, \ \forall\, i \in [n]$ is not feasible.

To deal with this issue, we relax the constraints to ℓ_1 **norm soft-margin** constraints:

$$y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_i) + b) \geq 1 - \xi_i, \ \forall\, i \in [n]$$

where we introduce **slack variables** $\xi_i \geq 0$.

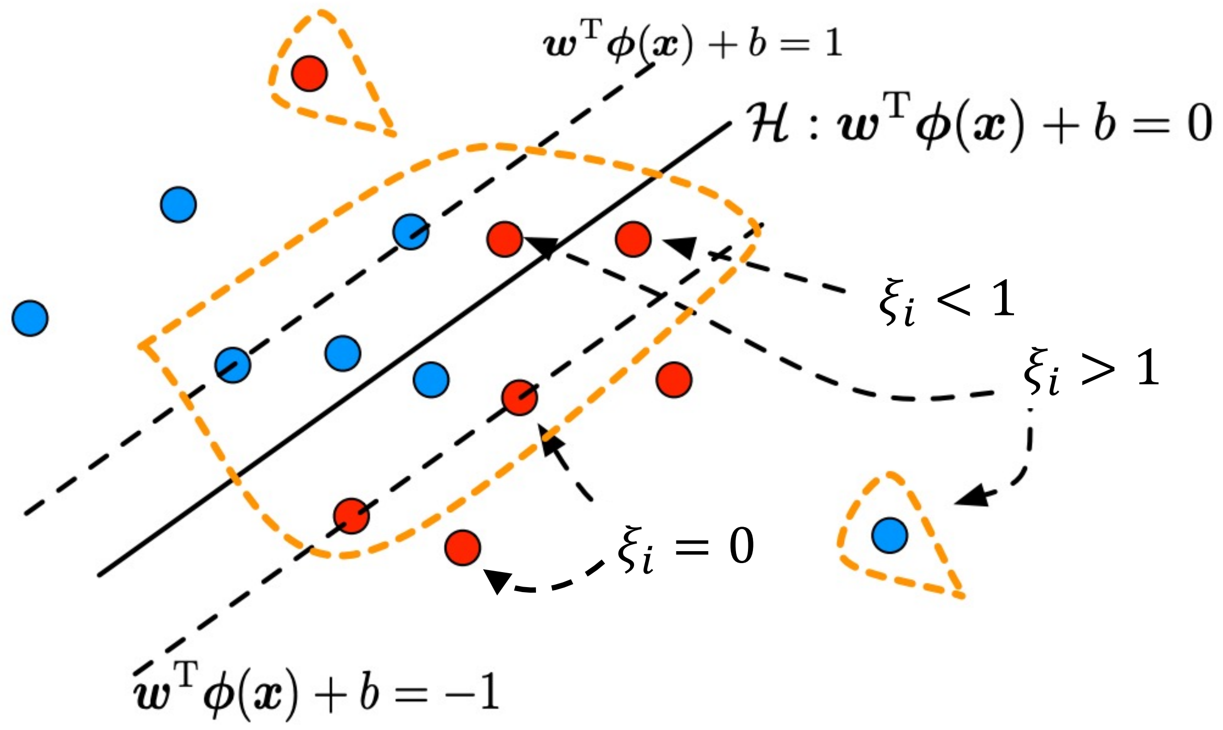
1.5 SVM General Primal Formulation

We want ξ_i to be as small as possible. The objective becomes

$$\begin{aligned} \min_{\boldsymbol{w}, b, \{\xi_i\}} \quad & \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_i \xi_i \\ \text{s.t.} \quad & y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_i) + b) \geq 1 - \xi_i, \quad \forall\, i \in [n] \\ & \xi_i \geq 0, \quad \forall\, i \in [n] \end{aligned}$$

where C is a hyperparameter to balance the two goals.

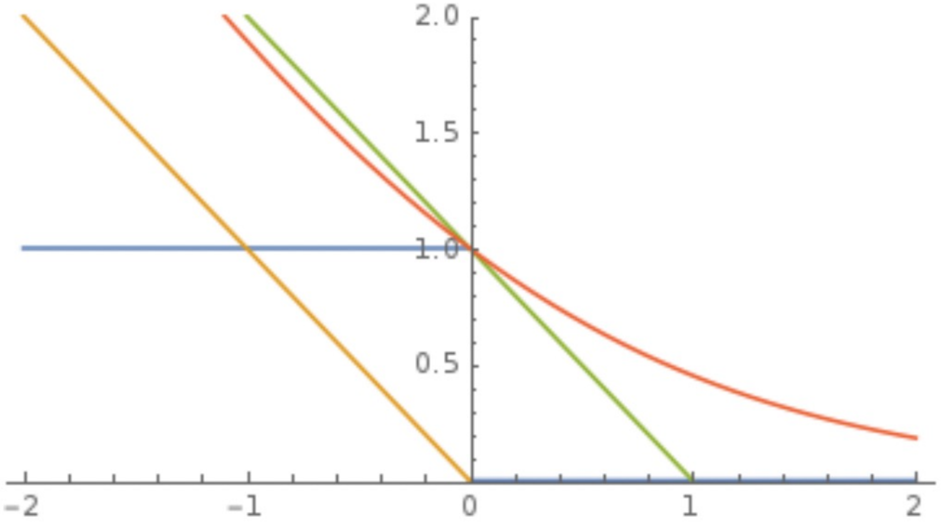
1.6 Understanding the slack conditions



- when $\xi_i^* = 0$, point is classified correctly and satisfies large margin constraint.
- when $\xi_i^* < 1$, point is classified correctly but does not satisfy large margin constraint.
- when $\xi_i^* > 1$, point is misclassified.

1.7 Primal formulation: Another view

In one sentence: **linear model with ℓ_2 regularized hinge loss**. Recall:



- perceptron loss $\ell_{\text{perceptron}}(z) = \max\{0, -z\} \to$ Perceptron
- logistic loss $\ell_{\text{logistic}}(z) = \log(1 + \exp(-z)) \to$ logistic regression
- hinge loss $\ell_{\text{hinge}}(z) = \max\{0, 1 - z\} \to$ **SVM**

1.7 Primal formulation: Another view

For a linear model $(\boldsymbol{w}, b)$, this means

$$\min_{\boldsymbol{w},b} \sum_i \max\left\{0, 1 - y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_i) + b)\right\} + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2$$

- recall $y_i \in \{-1, +1\}$
- a nonlinear mapping ϕ is applied
- the bias/intercept term b is used explicitly (why is this done?)

What is the relation between this formulation and the one which we just saw before?

1.7 Equivalent forms

The formulation

$$
\begin{aligned}
\min_{\boldsymbol{w}, b, \{\xi_i\}} \quad & C \sum_i \xi_i + \frac{1}{2} \|\boldsymbol{w}\|_2^2 \\
\text{s.t.} \quad & 1 - y_i(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}_i) + b) \leq \xi_i, \quad \forall\, i \in [n] \\
& \xi_i \geq 0, \quad \forall\, i \in [n]
\end{aligned}
$$

is equivalent to

$$
\begin{aligned}
\min_{\boldsymbol{w}, b, \{\xi_i\}} \quad & C \sum_i \xi_i + \frac{1}{2} \|\boldsymbol{w}\|_2^2 \\
\text{s.t.} \quad & \max\left\{0, 1 - y_i(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}_i) + b)\right\} = \xi_i, \quad \forall\, i \in [n]
\end{aligned}
$$

1.7 Equivalent forms

$$\min_{\boldsymbol{w},b,\{\xi_i\}} \quad C\sum_i \xi_i + \frac{1}{2}\|\boldsymbol{w}\|_2^2$$

$$\text{s.t.} \quad \max\left\{0, 1 - y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_i) + b)\right\} = \xi_i, \quad \forall\, i \in [n]$$

is equivalent to

$$\min_{\boldsymbol{w},b} C\sum_i \max\left\{0, 1 - y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_i) + b)\right\} + \frac{1}{2}\|\boldsymbol{w}\|_2^2$$

and

$$\min_{\boldsymbol{w},b} \sum_i \max\left\{0, 1 - y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_i) + b)\right\} + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2$$

with $\lambda = 1/C$. *This is exactly minimizing ℓ_2 regularized hinge loss!*

1.8 Optimization

$$
\begin{aligned}
\min_{\boldsymbol{w}, b, \{\xi_i\}} \quad & C \sum_i \xi_i + \frac{1}{2} \|\boldsymbol{w}\|_2^2 \\
\text{s.t.} \quad & y_i(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}_i) + b) \geq 1 - \xi_i, \quad \forall\, i \in [n] \\
& \xi_i \geq 0, \quad \forall\, i \in [n].
\end{aligned}
$$

- it is a convex (in fact, a **quadratic**) problem
- thus can apply any convex optimization algorithms, e.g. SGD
- there are **more specialized and efficient** algorithms
- but usually we apply kernel trick, which requires solving the *dual problem*

SVMs: Dual formulation & Kernel trick

How did we show this for regularized least squares?

By setting the gradient of $F(\boldsymbol{w}) = \|\boldsymbol{\Phi}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{w}\|_2^2$ to be $\mathbf{0}$:

$$\boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{\Phi}\boldsymbol{w}^* - \boldsymbol{y}) + \lambda\boldsymbol{w}^* = \mathbf{0}$$

we know

$$\boldsymbol{w}^* = \frac{1}{\lambda}\boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{w}^*) = \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\alpha} = \sum_{i=1}^{n} \alpha_i \boldsymbol{\phi}(\boldsymbol{x}_i)$$

Thus the least square solution is **a linear combination of features of the datapoints**!

2.1 Kernelizing SVM

We can also geometrically understand why $\boldsymbol{w}^*$ should lie in the span of the data:

$$\begin{aligned} \min_{\boldsymbol{w}, b} \quad & \frac{1}{2}\|\boldsymbol{w}\|_2^2 \\ \text{s.t.} \quad & y_j(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_j) + b) \geq 1, \;\; \forall\, j \in [n]. \end{aligned}$$

2.2 SVM: Dual form for separable case

With some optimization theory (Lagrange duality, not covered in this class), we can show this is equivalent to,

$$
\begin{aligned}
\max_{\{\alpha_i\}} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j \boldsymbol{\phi}(\boldsymbol{x}_i)^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}_j) \\
\text{s.t.} \quad & \sum_{i=1}^{n} \alpha_i y_i = 0 \quad \text{and} \quad \alpha_i \geq 0, \quad \forall\, i \in [n]
\end{aligned}
$$

2.2 SVM: Dual form for separable case

Using the kernel function k for the mapping $\boldsymbol{\phi}$, we can kernelize this!

$$\max_{\{\alpha_i\}} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j k(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \quad \text{and} \quad \alpha_i \geq 0, \quad \forall\, i \in [n]$$

No need to compute $\phi(\boldsymbol{x})$. This is also a **quadratic program** and many efficient optimization algorithms exist.

2.3 SVM: Dual form for the general case

For the primal for the general (non-separable) case:

$$\begin{aligned} \min_{\boldsymbol{w},b,\{\xi_i\}} \quad & C\sum_i \xi_i + \frac{1}{2}\|\boldsymbol{w}\|_2^2 \\ \text{s.t.} \quad & y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_i) + b) \geq 1 - \xi_i, \quad \forall\, i \in [n] \\ & \xi_i \geq 0, \quad \forall\, i \in [n]. \end{aligned}$$

The dual is very similar,

$$\begin{aligned} \max_{\{\alpha_i\}} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i,j} y_i y_j \alpha_i \alpha_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ \text{s.t.} \quad & \sum_{i=1}^n \alpha_i y_i = 0 \quad \text{and} \quad 0 \leq \alpha_i \leq C, \quad \forall\, i \in [n]. \end{aligned}$$

2.4 Prediction using SVM

How do we predict given the solution $\{\alpha_i^*\}$ to the dual optimization problem?

Remember that,

$$\boldsymbol{w}^* = \sum_i \alpha_i^* y_i \boldsymbol{\phi}(\boldsymbol{x}_i) = \sum_{i:\alpha_i^*>0} \alpha_i^* y_i \boldsymbol{\phi}(\boldsymbol{x}_i)$$

A point with $\alpha_i^* > 0$ is called a "**support vector**". Hence the name SVM.

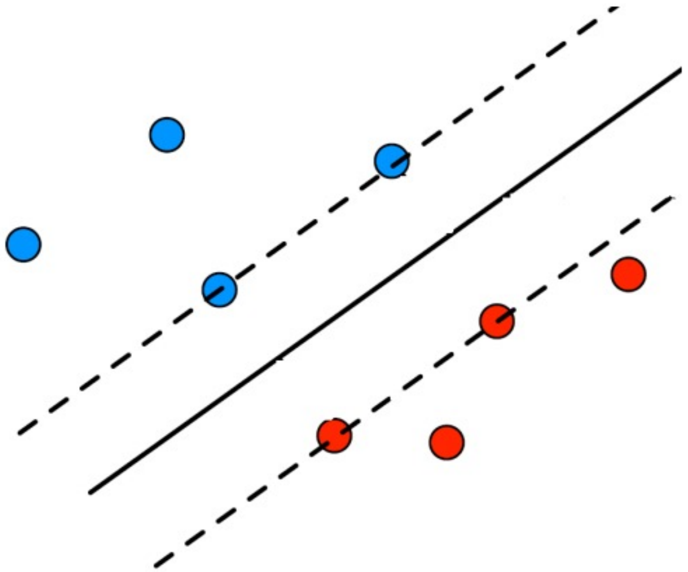
To make a prediction on any datapoint $\boldsymbol{x}$,

$$\begin{aligned}\text{sign}\left(\boldsymbol{w}^{*\mathrm{T}}\phi(\boldsymbol{x}) - b^*\right) &= \text{sign}\left(\sum_{i:\alpha_i^*>0} \alpha_i^* y_i \boldsymbol{\phi}(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}) - b^*\right) \\ &= \text{sign}\left(\sum_{i:\alpha_i^*>0} \alpha_i^* y_i k(\boldsymbol{x}_i, \boldsymbol{x}) - b^*\right).\end{aligned}$$

All we need now is to identify b^*.

2.5 Bias term b^*

First, let's consider the separable case:



It can be shown (we will not cover in class), that in the separable case the support vectors lie on the margin.

2.5 Bias term b^*

General (non-separable case):

For any support vector $\boldsymbol{\phi}(\boldsymbol{x}_i)$ with $0 < \alpha_i^* < C$, it can be shown that $1 = y_i({\boldsymbol{w}^*}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_i) + b^*)$ (i.e. that support vector lies on the margin). Therefore, as before,

$$b^* = y_i - {\boldsymbol{w}^*}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_i) = y_i - \sum_{j=1}^{n} \alpha_j^* y_j k(\boldsymbol{x}_j, \boldsymbol{x}_i).$$

In practice, often *average* over all i with $0 < \alpha_i^* < C$ to stabilize computation.

With $\boldsymbol{\alpha}^*$ and b^* in hand, we can make a prediction on any datapoint $\boldsymbol{x}$,

$$\operatorname{sign}\left({\boldsymbol{w}^*}^{\mathrm{T}}\phi(\boldsymbol{x}) - b^*\right) = \operatorname{sign}\left(\sum_{i:\alpha_i^*>0} \alpha_i^* y_i k(\boldsymbol{x}_i, \boldsymbol{x}) - b^*\right).$$

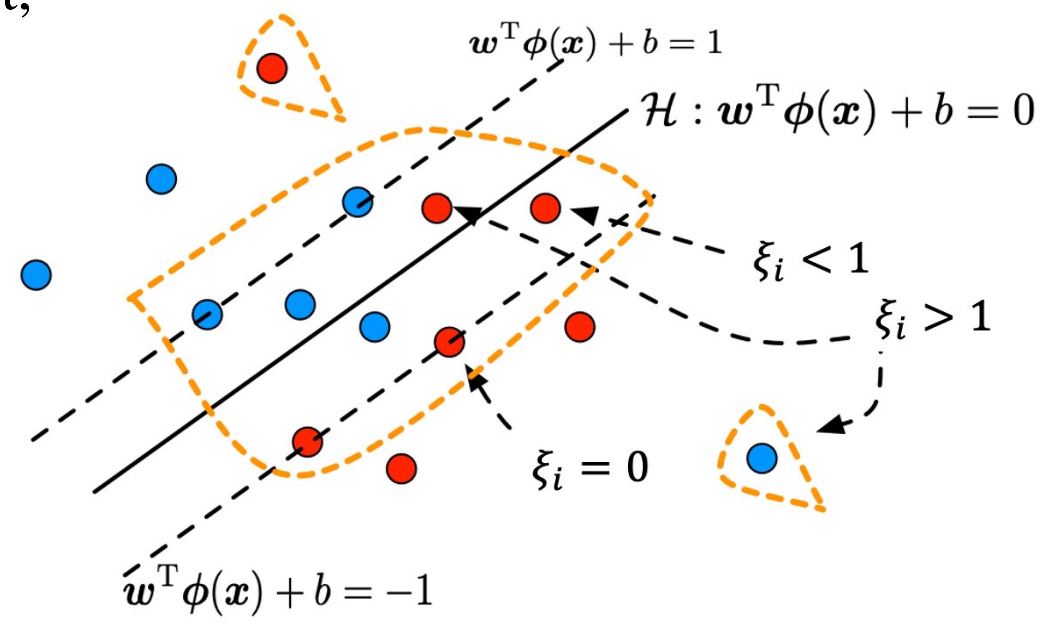
SVMs: Understanding them further

3.1 Understanding support vectors

Support vectors are $\boldsymbol{\phi}(\boldsymbol{x}_i)$ such that $\alpha_i^* > 0$.

They are the set of points which satisfy one of the following:
(1) they are tight with respect to the large margin contraint,
(2) they do not satisfy the large margin contraint,
(3) they are misclassified.

- when $\xi_i^* = 0$, $y_i({\boldsymbol{w}^*}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_i) + b^*) = 1$,
  and thus the point is $1/\|\boldsymbol{w}^*\|_2$ away from the hyperplane.
- when $\xi_i^* < 1$, the point is classified correctly
  but does not satisfy the large margin constraint.
- when $\xi_i^* > 1$, the point is misclassified.



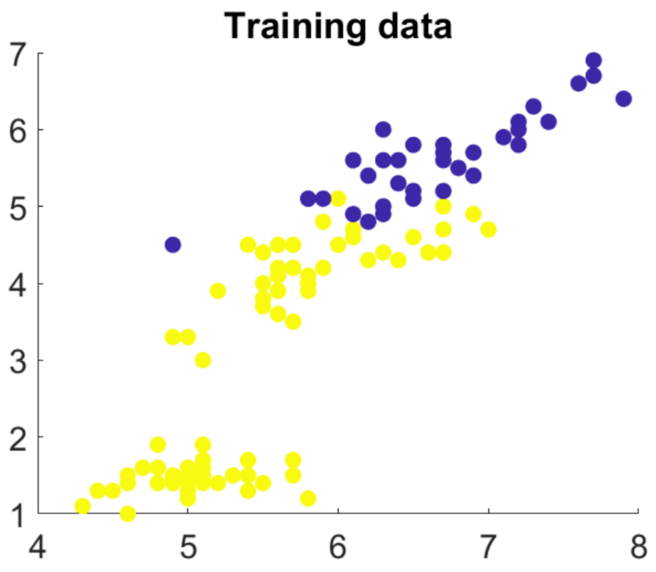
Support vectors (circled with the orange line) are the only points that matter!

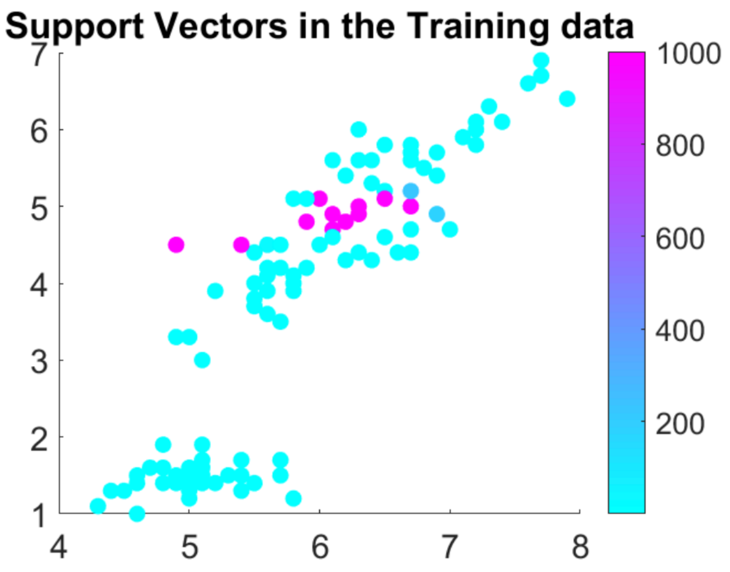
3.1 Understanding support vectors

One potential drawback of kernel methods: **non-parametric**, need to potentially keep all the training points.

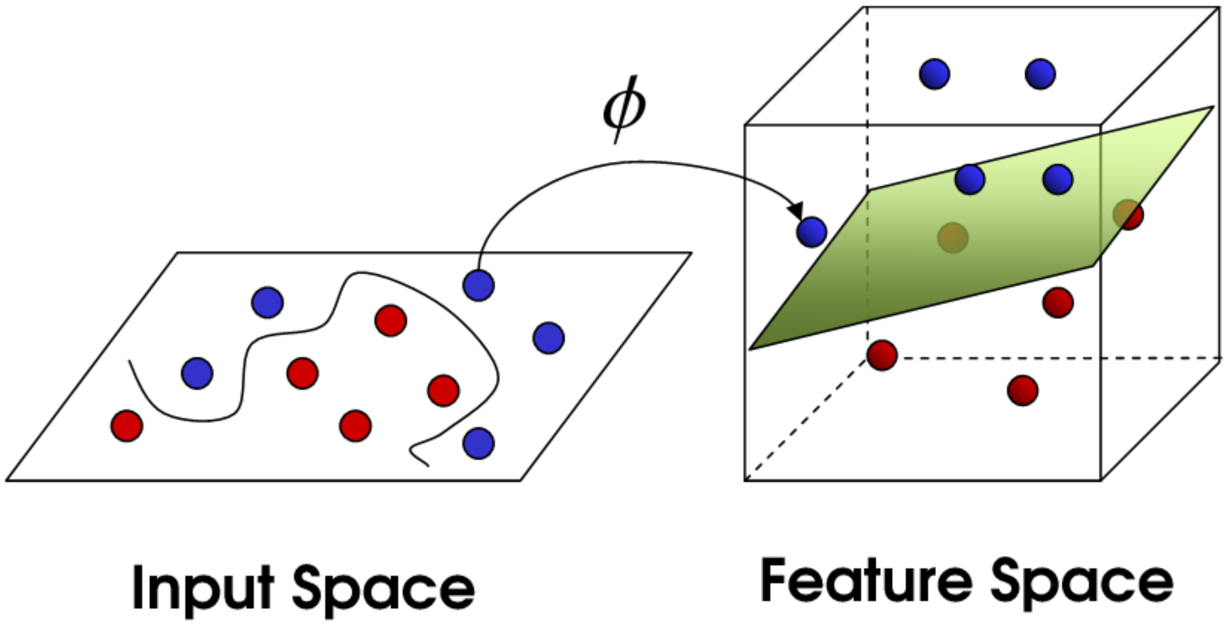
$$\text{sign}\left(\boldsymbol{w}^{*\mathrm{T}}\phi(\boldsymbol{x}) - b^*\right) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i^* y_i k(\boldsymbol{x}_i, \boldsymbol{x}) - b^*\right).$$

For SVM though, very often #support vectors $= \left|\{i : \alpha_i^* > 0\}\right| \ll n$.



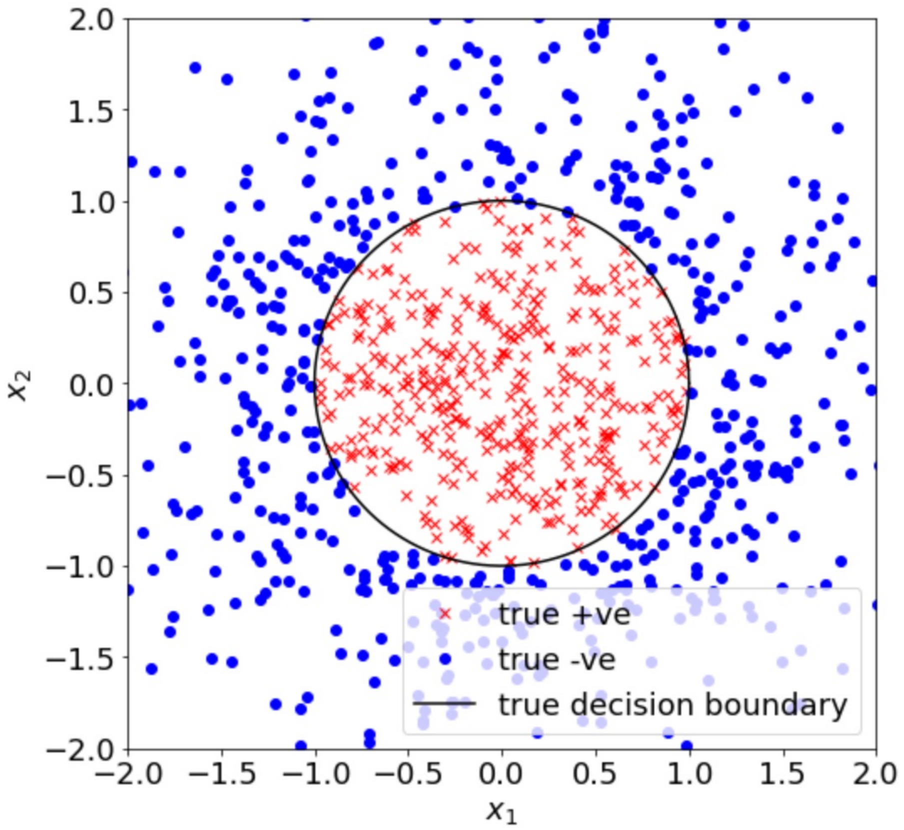


3.2 Examining the effect of kernels



Data may become linearly separable when lifted to the high-dimensional feature space!

Polynomial kernel: example



Switch to Colab

Gaussian kernel: example

Gaussian kernel or Radial basis function (RBF) kernel

$$k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2}{2\sigma^2}\right)$$

for some $\sigma > 0$. This is also parameterized as,

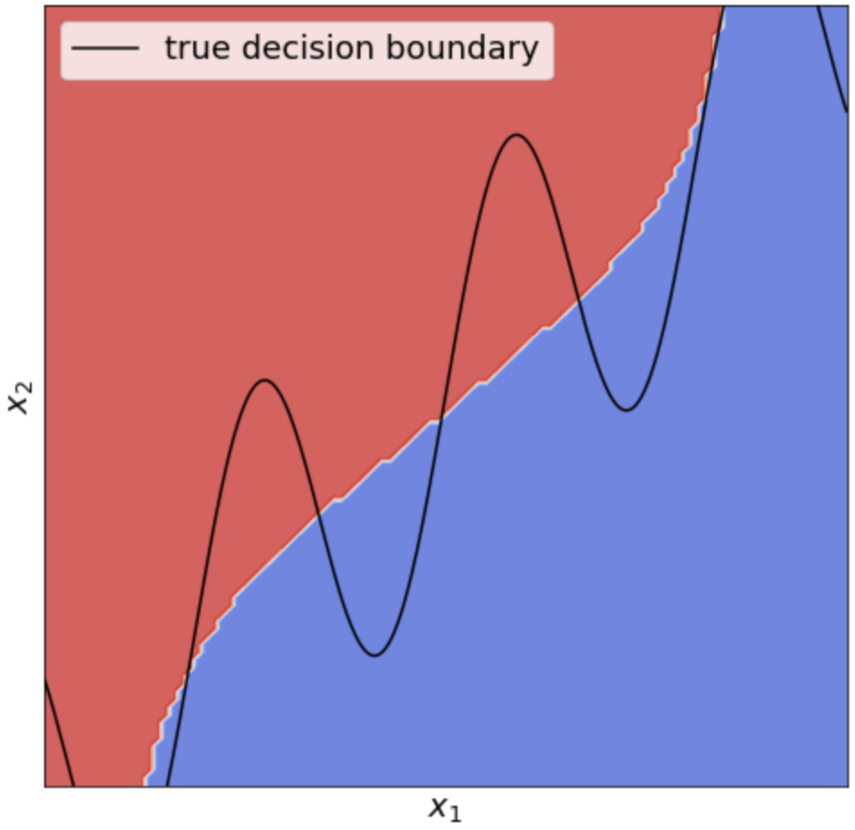
$$k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\gamma\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2\right)$$

for some $\gamma > 0$.

What does the decision boundary look like?
What is the effect of γ?

Note that the prediction is of the form

$$\operatorname{sign}\left(\boldsymbol{w}^{*\mathrm{T}}\phi(\boldsymbol{x}) - b^*\right) = \operatorname{sign}\left(\sum_{i:\alpha_i^*>0} \alpha_i^* y_i k(\boldsymbol{x}_i, \boldsymbol{x}) - b^*\right).$$



Switch to Colab

SVM: Summary of mathematical forms

SVM: **max-margin linear classifier**

Primal (equivalent to minimizing ℓ_2 regularized hinge loss):

$$\begin{aligned} \min_{\boldsymbol{w}, b, \{\xi_i\}} \quad & C \sum_i \xi_i + \frac{1}{2}\|\boldsymbol{w}\|_2^2 \\ \text{s.t.} \quad & y_i(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}_i) + b) \geq 1 - \xi_i, \quad \forall\, i \in [n] \\ & \xi_i \geq 0, \quad \forall\, i \in [n]. \end{aligned}$$

Dual (kernelizable, reveals what training points are support vectors):

$$\begin{aligned} \max_{\{\alpha_i\}} \quad & \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j \boldsymbol{\phi}(\boldsymbol{x}_i)^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}_j) \\ \text{s.t.} \quad & \sum_i \alpha_i y_i = 0 \quad \text{and} \quad 0 \leq \alpha_i \leq C, \quad \forall\, i \in [n]. \end{aligned}$$

Multiclass classification

4.1 Setup

Recall the setup:

- input (feature vector): $\boldsymbol{x} \in \mathbb{R}^d$
- output (label): $y \in [\mathsf{C}] = \{1, 2, \cdots, \mathsf{C}\}$
- goal: learn a mapping $f : \mathbb{R}^d \to [\mathsf{C}]$

Examples:

- recognizing digits ($\mathsf{C} = 10$) or letters ($\mathsf{C} = 26$ or 52)
- predicting weather: sunny, cloudy, rainy, etc
- predicting image category: ImageNet dataset ($\mathsf{C} \approx 20K$)

4.2 Linear models: Binary to multiclass

Step 1: *What should a linear model look like for multiclass tasks?*

Note: a linear model for binary tasks (switching from $\{-1, +1\}$ to $\{1, 2\}$)

$$f(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} \geq 0 \\ 2 & \text{if } \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} < 0 \end{cases}$$

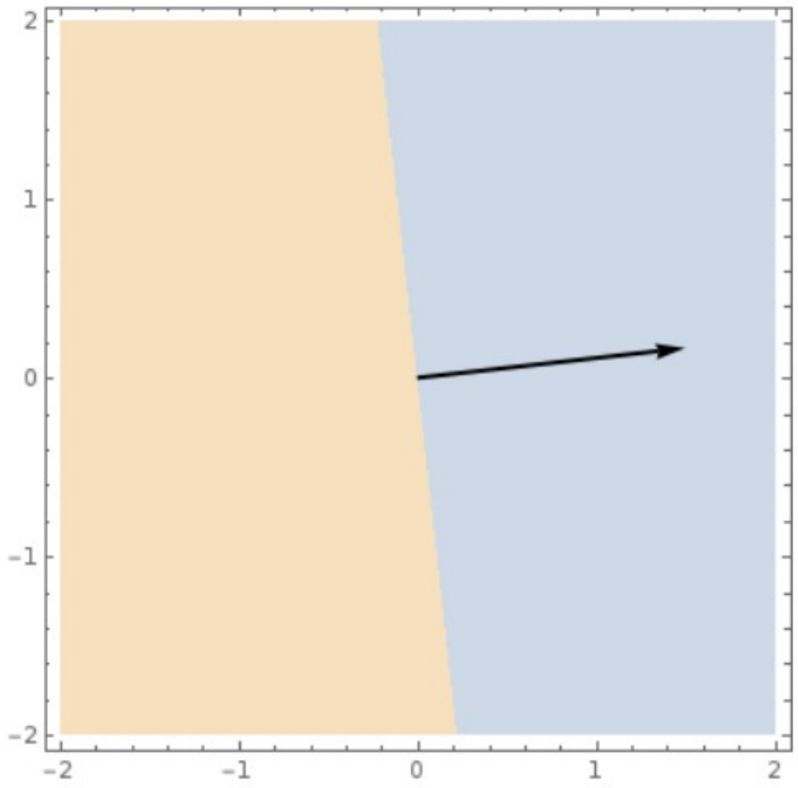
can be written as

$$\begin{aligned} f(\boldsymbol{x}) &= \begin{cases} 1 & \text{if } \boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x} \geq \boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{x} \\ 2 & \text{if } \boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{x} > \boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x} \end{cases} \\ &= \operatorname*{argmax}_{k \in \{1,2\}} \boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x} \end{aligned}$$

for any $\boldsymbol{w}_1, \boldsymbol{w}_2$ s.t. $\boldsymbol{w} = \boldsymbol{w}_1 - \boldsymbol{w}_2$

Think of $\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x}$ as **a score for class k**.

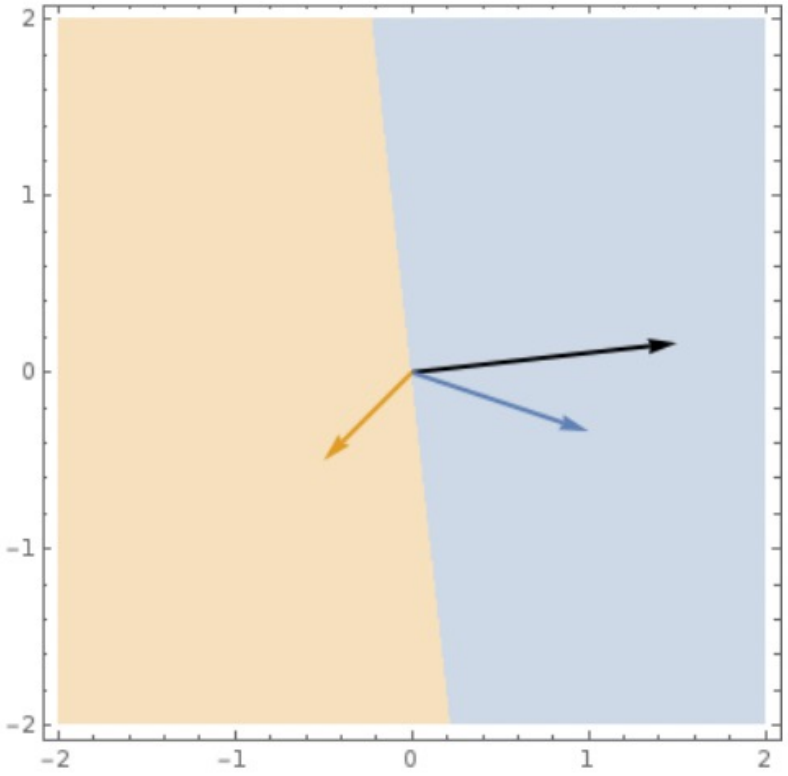
4.2 Linear models: Binary to multiclass



$$\boldsymbol{w} = \left(\tfrac{3}{2}, \tfrac{1}{6}\right)$$

- Blue class: $\{\boldsymbol{x} : \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} \geq 0\}$
- Orange class: $\{\boldsymbol{x} : \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} < 0\}$

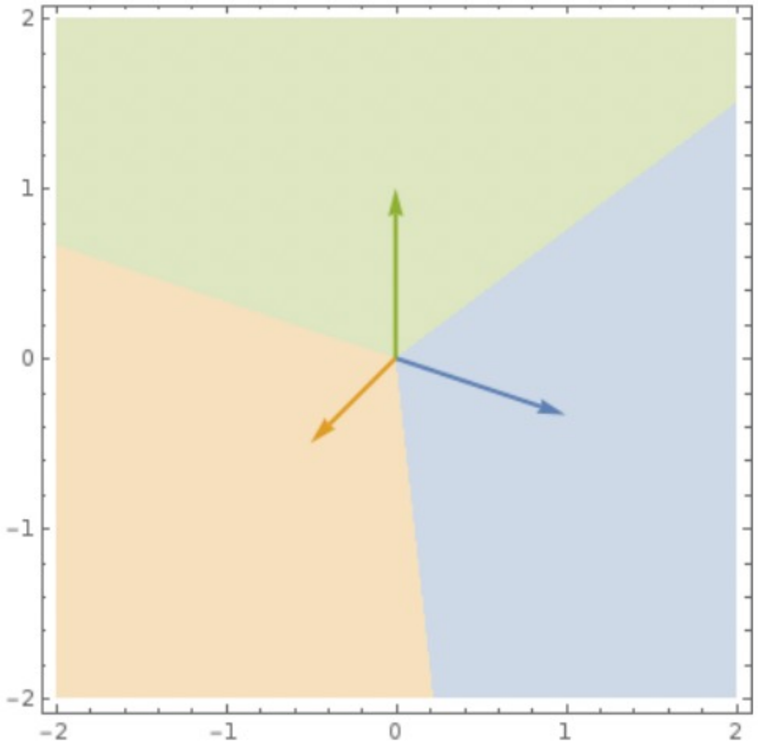
4.2 Linear models: Binary to multiclass



$$\boldsymbol{w} = \left(\tfrac{3}{2}, \tfrac{1}{6}\right) = \boldsymbol{w}_1 - \boldsymbol{w}_2$$
$$\boldsymbol{w}_1 = \left(1, -\tfrac{1}{3}\right)$$
$$\boldsymbol{w}_2 = \left(-\tfrac{1}{2}, -\tfrac{1}{2}\right)$$

- Blue class:
  $\{\boldsymbol{x} : 1 = \mathrm{argmax}_k\, \boldsymbol{w}_k^{\mathrm{T}} \boldsymbol{x}\}$
- Orange class:
  $\{\boldsymbol{x} : 2 = \mathrm{argmax}_k\, \boldsymbol{w}_k^{\mathrm{T}} \boldsymbol{x}\}$

4.2 Linear models: Binary to multiclass



$\boldsymbol{w}_1 = (1, -\frac{1}{3})$
$\boldsymbol{w}_2 = (-\frac{1}{2}, -\frac{1}{2})$
$\boldsymbol{w}_3 = (0, 1)$

- Blue class:
  $\{\boldsymbol{x} : 1 = \mathrm{argmax}_k\, \boldsymbol{w}_k^{\mathrm{T}} \boldsymbol{x}\}$
- Orange class:
  $\{\boldsymbol{x} : 2 = \mathrm{argmax}_k\, \boldsymbol{w}_k^{\mathrm{T}} \boldsymbol{x}\}$
- Green class:
  $\{\boldsymbol{x} : 3 = \mathrm{argmax}_k\, \boldsymbol{w}_k^{\mathrm{T}} \boldsymbol{x}\}$

4.3 Function class: Linear models for multiclass classification

$$\mathcal{F} = \left\{ f(\boldsymbol{x}) = \operatorname*{argmax}_{k \in [\mathsf{C}]} \ \boldsymbol{w}_k^{\mathrm{T}} \boldsymbol{x} \mid \boldsymbol{w}_1, \ldots, \boldsymbol{w}_{\mathsf{C}} \in \mathbb{R}^d \right\}$$

$$= \left\{ f(\boldsymbol{x}) = \operatorname*{argmax}_{k \in [\mathsf{C}]} (\boldsymbol{W}\boldsymbol{x})_k \mid \boldsymbol{W} \in \mathbb{R}^{\mathsf{C} \times d} \right\}$$

Next, lets try to generalize the loss functions. Focus on the logistic loss today.

4.4 Multinomial logistic regression: a probabilistic view

Observe: for binary logistic regression, with $\boldsymbol{w} = \boldsymbol{w}_1 - \boldsymbol{w}_2$:

$$\Pr(y = 1 \mid \boldsymbol{x}; \boldsymbol{w}) = \sigma(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}}} = \frac{e^{\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x}}}{e^{\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x}} + e^{\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{x}}} \propto e^{\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x}}$$

Naturally, for multiclass:

$$\Pr(y = k \mid \boldsymbol{x}; \boldsymbol{W}) = \frac{e^{\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x}}}{\sum_{k \in [\mathsf{C}]} e^{\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x}}} \propto e^{\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x}}$$

This is called the *softmax function*.

4.5 Let's find the **MLE**

Maximize probability of seeing labels $y_1, \ldots, y_n$ given $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$

$$P(\boldsymbol{W}) = \prod_{i=1}^{n} \Pr(y_i \mid \boldsymbol{x}_i; \boldsymbol{W}) = \prod_{i=1}^{n} \frac{e^{\boldsymbol{w}_{y_i}^{\mathrm{T}} \boldsymbol{x}_i}}{\sum_{k \in [\mathsf{C}]} e^{\boldsymbol{w}_k^{\mathrm{T}} \boldsymbol{x}_i}}$$

By taking **negative log**, this is equivalent to minimizing

$$F(\boldsymbol{W}) = \sum_{i=1}^{n} \ln \left(\frac{\sum_{k \in [\mathsf{C}]} e^{\boldsymbol{w}_k^{\mathrm{T}} \boldsymbol{x}_i}}{e^{\boldsymbol{w}_{y_i}^{\mathrm{T}} \boldsymbol{x}_i}} \right) = \sum_{i=1}^{n} \ln \left(1 + \sum_{k \neq y_i} e^{(\boldsymbol{w}_k - \boldsymbol{w}_{y_i})^{\mathrm{T}} \boldsymbol{x}_i} \right)$$

This is the *multiclass logistic loss*. It is an upper-bound on the 0-1 misclassification loss:

$$\mathbb{I}[f(\boldsymbol{x}) \neq y] \leq \log_2 \left(1 + \sum_{k \neq y} e^{(\boldsymbol{w}_k - \boldsymbol{w}_y)^{\mathrm{T}} \boldsymbol{x}} \right)$$

When $\mathsf{C} = 2$, multiclass logistic loss is the same as binary logistic loss (let's verify).

Relating binary and multiclass logistic loss

4.6 Next, optimization

Apply **SGD**: what is the gradient of

$$F(\boldsymbol{W}) = \ln\left(1 + \sum_{k \neq y_i} e^{(\boldsymbol{w}_k - \boldsymbol{w}_{y_i})^{\mathrm{T}}\boldsymbol{x}_i}\right)?$$

It's a C $\times$ d matrix. Let's focus on the k-th row:

If $k \neq y_i$:

$$\nabla_{\boldsymbol{w}_k^{\mathrm{T}}} F(\boldsymbol{W}) = \frac{e^{(\boldsymbol{w}_k - \boldsymbol{w}_{y_i})^{\mathrm{T}}\boldsymbol{x}_i}}{1 + \sum_{k \neq y_i} e^{(\boldsymbol{w}_k - \boldsymbol{w}_{y_i})^{\mathrm{T}}\boldsymbol{x}_i}} \boldsymbol{x}_i^{\mathrm{T}} = \frac{e^{\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x}_i}}{e^{\boldsymbol{w}_{y_i}^{\mathrm{T}}\boldsymbol{x}_i} + \sum_{k \neq y_i} e^{\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x}_i}} \boldsymbol{x}_i^{\mathrm{T}} = \Pr(y = k \mid \boldsymbol{x}_i; \boldsymbol{W})\boldsymbol{x}_i^{\mathrm{T}}$$

else:

$$\nabla_{\boldsymbol{w}_k^{\mathrm{T}}} F(\boldsymbol{W}) = \frac{-\left(\sum_{k \neq y_i} e^{(\boldsymbol{w}_k - \boldsymbol{w}_{y_i})^{\mathrm{T}}\boldsymbol{x}_i}\right)}{1 + \sum_{k \neq y_i} e^{(\boldsymbol{w}_k - \boldsymbol{w}_{y_i})^{\mathrm{T}}\boldsymbol{x}_i}} \boldsymbol{x}_i^{\mathrm{T}} = \frac{-\left(\sum_{k \neq y_i} e^{\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x}_i}\right)}{e^{\boldsymbol{w}_{y_i}^{\mathrm{T}}\boldsymbol{x}_i} + \sum_{k \neq y_i} e^{\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x}_i}} \boldsymbol{x}_i^{\mathrm{T}} = \left(\Pr(y = y_i \mid \boldsymbol{x}_i; \boldsymbol{W}) - 1\right) \boldsymbol{x}_i^{\mathrm{T}}$$

SGD for multinomial logistic regression

Initialize $\boldsymbol{W} = \mathbf{0}$ (or randomly). Repeat:

1. pick $i \in [n]$ uniformly at random
2. update the parameters

$$\boldsymbol{W} \leftarrow \boldsymbol{W} - \eta \begin{pmatrix} \Pr(y = 1 \mid \boldsymbol{x}_i; \boldsymbol{W}) \\ \vdots \\ \Pr(y = y_i \mid \boldsymbol{x}_i; \boldsymbol{W}) - 1 \\ \vdots \\ \Pr(y = \mathsf{C} \mid \boldsymbol{x}_i; \boldsymbol{W}) \end{pmatrix} \boldsymbol{x}_i^{\mathrm{T}}$$

Think about why the algorithm makes sense intuitively.

4.7 Probabilities -> Prediction

Having learned $\boldsymbol{W}$, we can either

- make a *deterministic* prediction $\operatorname{argmax}_{k \in [\mathsf{C}]} \; \boldsymbol{w}_k^{\mathrm{T}} \boldsymbol{x}$
- make a *randomized* prediction according to $\Pr(k \mid \boldsymbol{x}; \boldsymbol{W}) \propto e^{\boldsymbol{w}_k^{\mathrm{T}} \boldsymbol{x}}$