CS-GY 6923: Lecture 9
Kernel Methods, Support Vector Machines

NYU Tandon School of Engineering, Prof. Christopher Musco

NON-LINEAR METHODS

- Many previous methods studied (regression, logistic regression) are considered linear methods. They make predictions based on $\langle \mathbf{x}, \boldsymbol{\beta} \rangle$ – i.e. based on weighted sums of features.
- In the next part of the course we move on to non-linear methods. Specifically, kernel methods and neural networks.
- Both are very closely related to feature transformations, which was a technique we already saw for using linear methods to learn non-linear concepts.

RECALL: k-NEAREST NEIGHBOR METHOD

k-NN algorithm: a simple but powerful baseline for classification.

Training data: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ where $y_1, \ldots, y_n \in \{1, \ldots, q\}$.

Classification algorithm:

Given new input $\mathbf{x}_{new}$,

SOME similarity function (sim)

- Compute $sim(\mathbf{x}_{new}, \mathbf{x}_1), \ldots, sim(\mathbf{x}_{new}, \mathbf{x}_n)$.[1]
- Let $\mathbf{x}_{j_1}, \ldots, \mathbf{x}_{j_k}$ be the training data vectors with highest similarity to $\mathbf{x}_{new}$.
- Predict y_{new} as $majority(y_{j_1}, \ldots, y_{j_k})$.

[1] $sim(\mathbf{x}_{new}, \mathbf{x}_i)$ is any chosen similarity function, like $1 - \|\mathbf{x}_{new} - \mathbf{x}_i\|_2$.

k-NEAREST NEIGHBOR METHOD

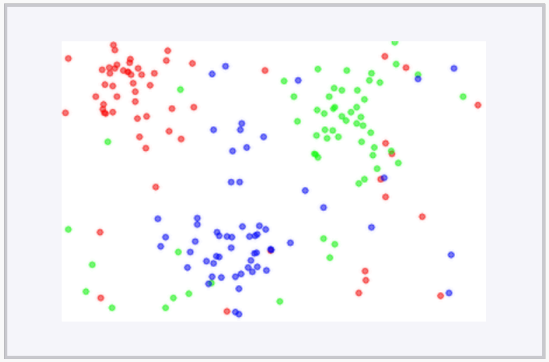


Fig. 1. The dataset.

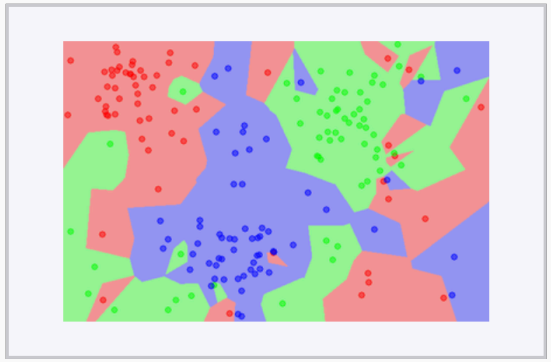


Fig. 2. The 1NN classification map.

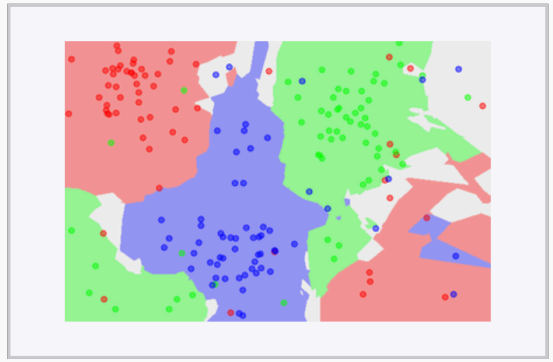


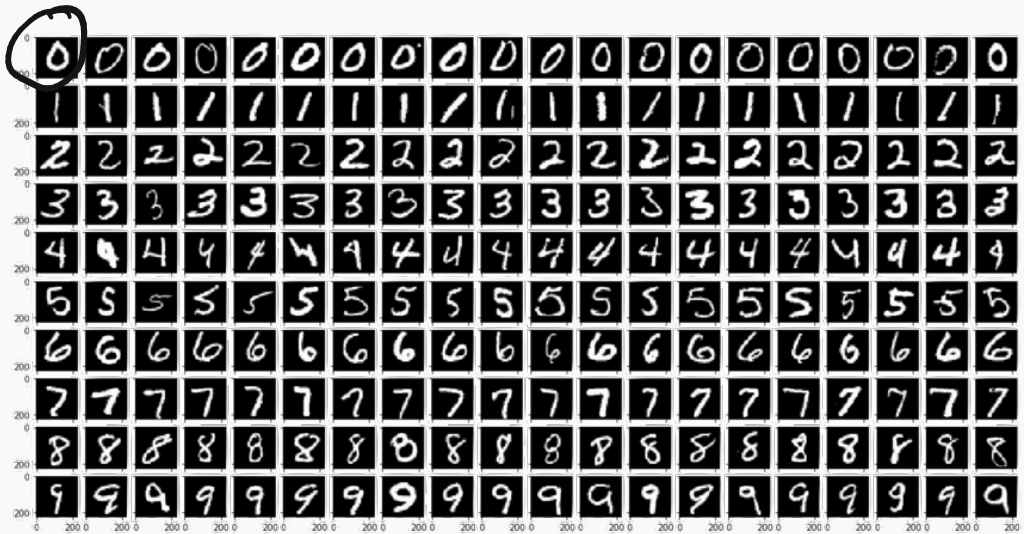
Fig. 3. The 5NN classification map.

- Smaller k, more complex classification function.
- Larger k, more robust to noisy labels.

Works remarkably well for many datasets.

MNIST IMAGE DATA

Especially good for large datasets with lots of repetition. Works well on MNIST for example. 95% Accuracy out-of-the-box.

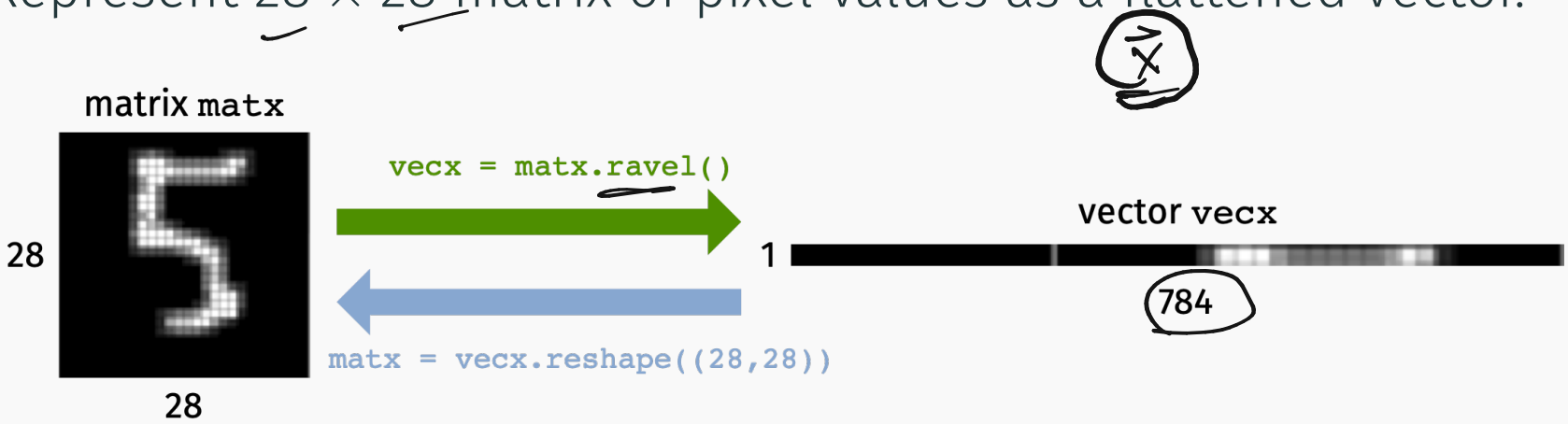


Let's look into this example a bit more...

MNIST IMAGE DATA

Each pixel is number from $[0, 1]$. 0 is black, 1 is white.

Represent 28×28 matrix of pixel values as a flattened vector.




```
xmat = np.array([[1,2,3],[4,5,6],[7,8,9]])
```

```
array([[1, 2, 3],
       [4, 5, 6],
       [7, 8, 9]])
```

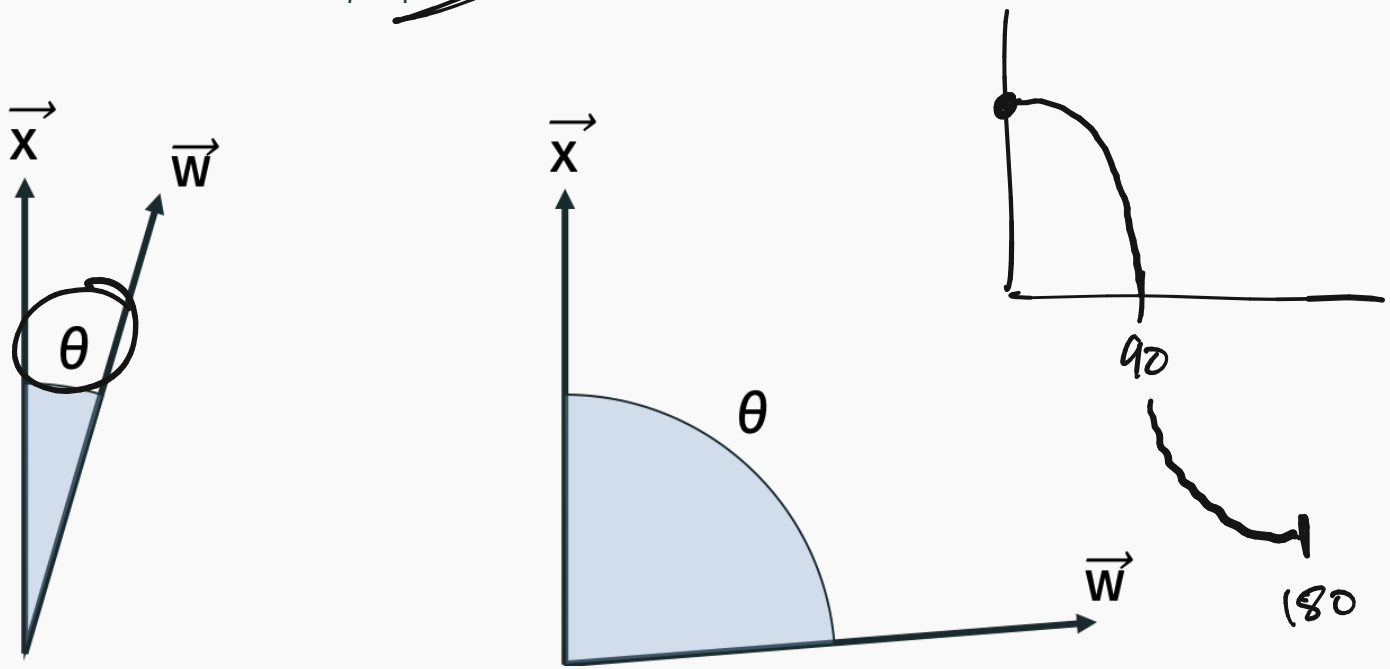
```
xvec = xmat.ravel()
```

```
array([1, 2, 3, 4, 5, 6, 7, 8, 9])
```

INNER PRODUCT SIMILARITY

Given data vectors $\mathbf{x}, \mathbf{w} \in \mathbb{R}^d$, the inner product $\langle \mathbf{x}, \mathbf{w} \rangle$ is a natural similarity measure.

$$\langle \mathbf{x}, \mathbf{w} \rangle = \sum_{i=1}^{d} x_i w_i = \cos(\theta) \|\mathbf{x}\|_2 \|\mathbf{w}\|_2.$$



Also called “cosine similarity”.

INNER PRODUCT SIMILARITY

$$(x-w)^T(x-w) = x^Tx + w^Tw - 2x^Tw$$

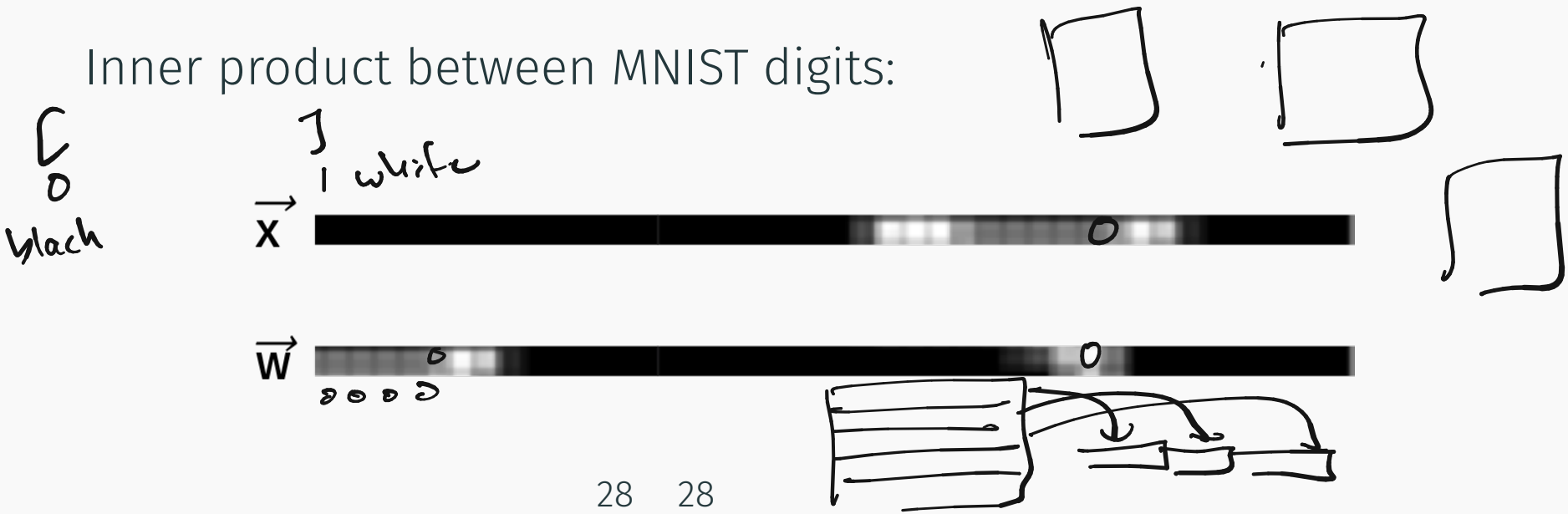
Connection to Euclidean (ℓ_2) Distance:

$$\|\mathbf{x} - \mathbf{w}\|_2^2 = \|\mathbf{x}\|_2^2 + \|\mathbf{w}\|_2^2 - 2\langle \mathbf{x}, \mathbf{w} \rangle$$

If all data vectors has the same norm, the pair of vectors with largest inner product is the pair with smallest Euclidean distance.

INNER PRODUCT FOR MNIST

Inner product between MNIST digits:

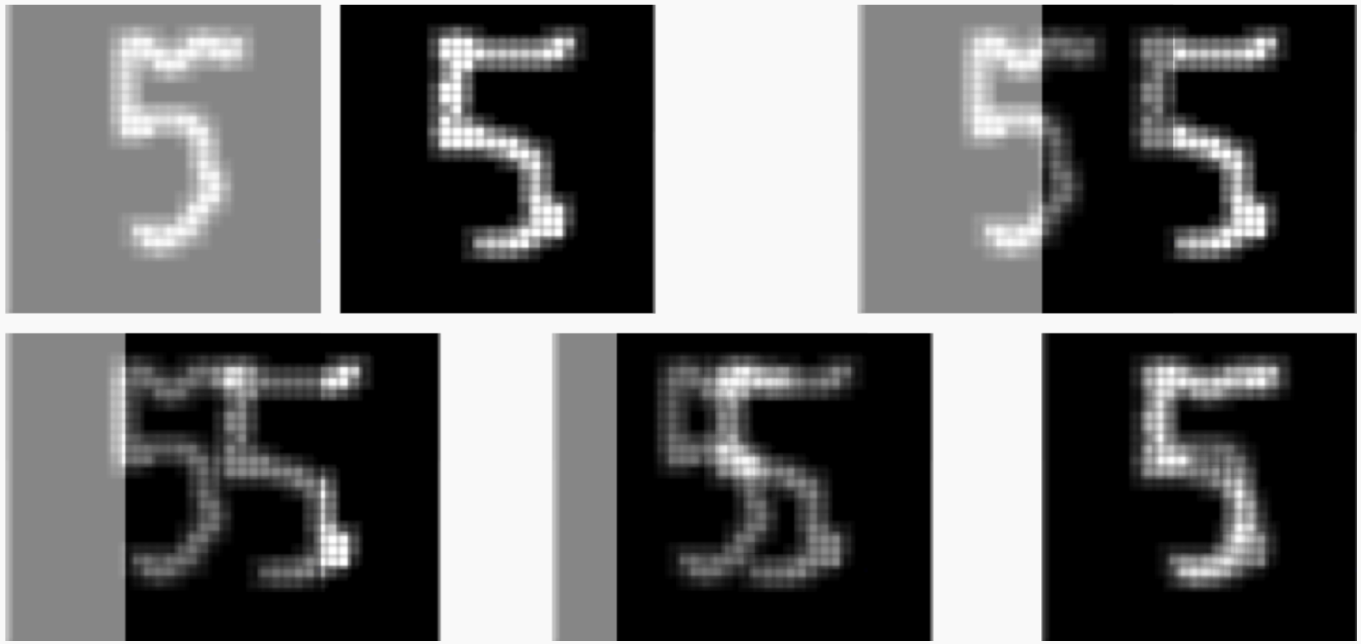


$$\langle \mathtt{x}, \mathtt{w} \rangle = \sum_{i=1}^{28} \sum_{j=1}^{28} \mathtt{matx}[i,j] \cdot \mathtt{matw}[i,j].$$

Inner product similarity is higher when the images have large pixel values (close to 1) in the same locations. I.e. when they have a lot of overlapping white/light gray pixels.

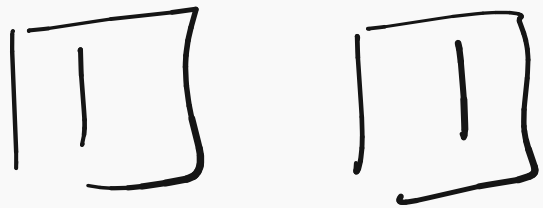
INNER PRODUCT FOR MNIST

Visualizing the inner product between two images:

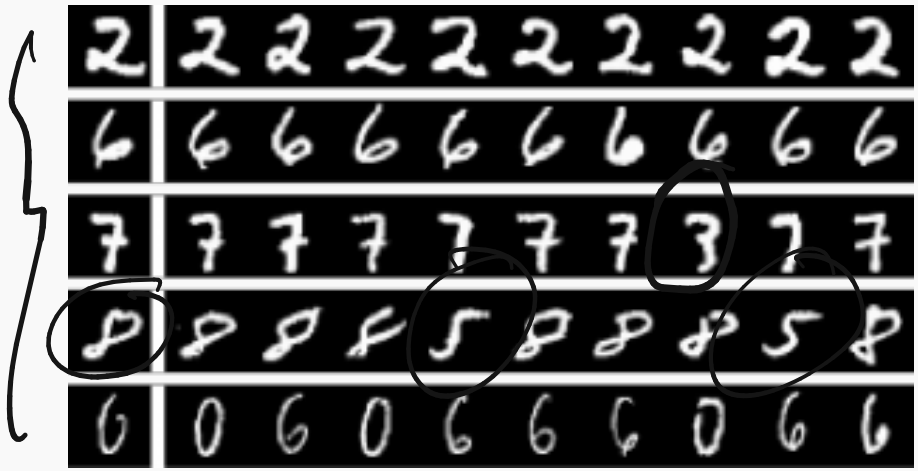


Images with high inner product have a lot of overlap.

K-NN ALGORITHM ON MNIST



Most similar images during k-nn search, $k = 9$:



$0, \ldots, q$

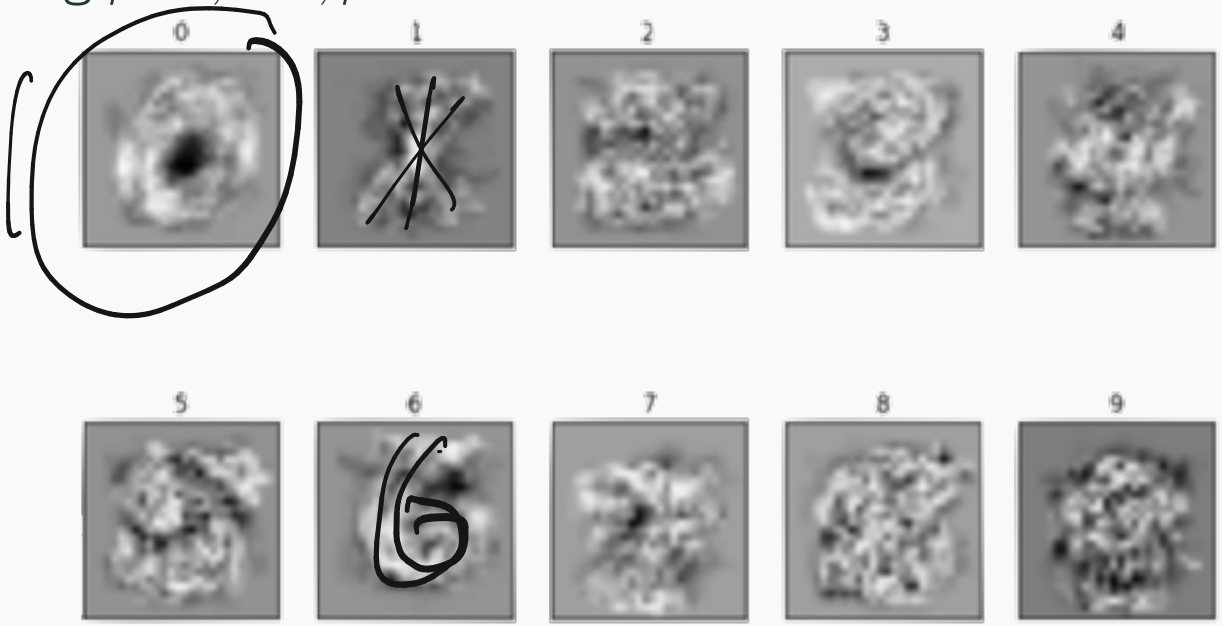
One-vs.-all or Multiclass Cross-entropy Classification with Logistic Regression:

- Learn q classifiers with parameters $\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \ldots, \boldsymbol{\beta}^{(q)}$.
- Given $\mathbf{x}_{new}$ compute $\langle \mathbf{x}_{new}, \boldsymbol{\beta}^{(1)} \rangle, \ldots, \langle \mathbf{x}_{new}, \boldsymbol{\beta}^{(q)} \rangle$
- Predict class $y_{new} = \arg\max_i \langle \mathbf{x}_{new}, \boldsymbol{\beta}^{(i)} \rangle$.

If each $\mathbf{x}$ is a vector with $28 \times 28 = 784$ entries than each $\boldsymbol{\beta}^{(i)}$ also has 784 entries. Each parameter vector can be viewed as a 28×28 image.

MATCHED FILTER

Visualizing $\beta^{(0)}, \ldots, \beta^{(9)}$:

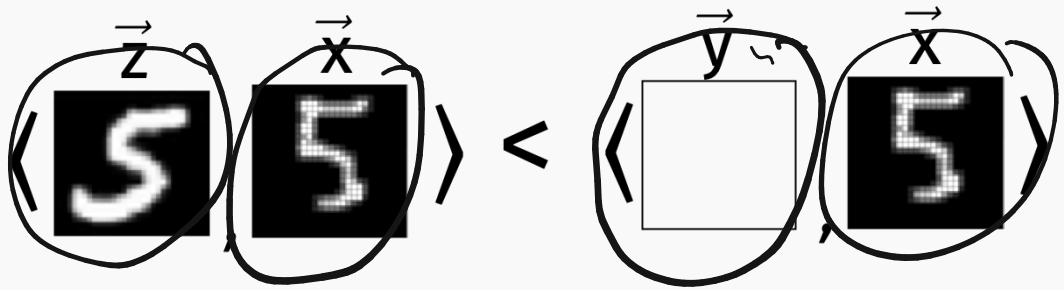


Logistic regression classification rule: For an input  , compute inner product similarity with all weight matrices and choose most similar one.

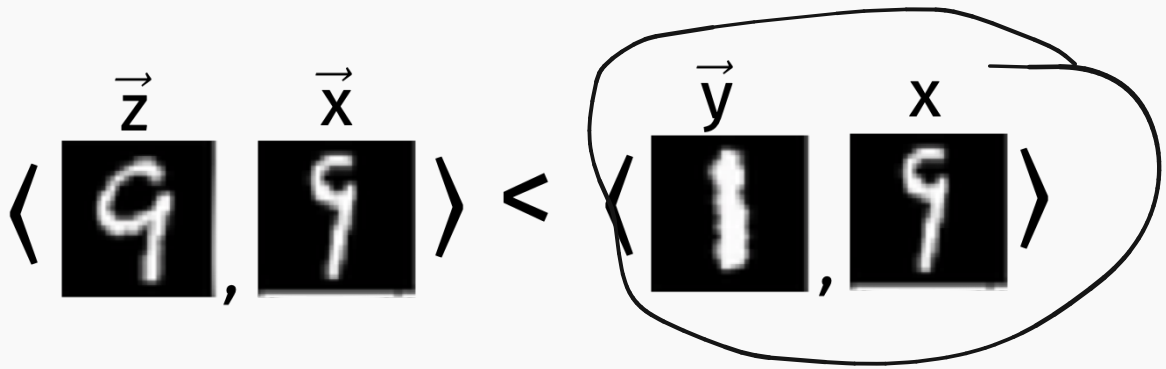
In contrast to k-NN, only needs to compute similarity with q items instead of n. Much faster classification.

DIVING INTO SIMILARITY

Often the inner product does not make sense as a similarity measure between data vectors. Here's an example (recall that smaller inner product means less similar):



But clearly the first image is more similar.



Here's a more realistic scenario.

KERNEL FUNCTIONS: ALTERNATIVE MEASURES OF SIMILARITY

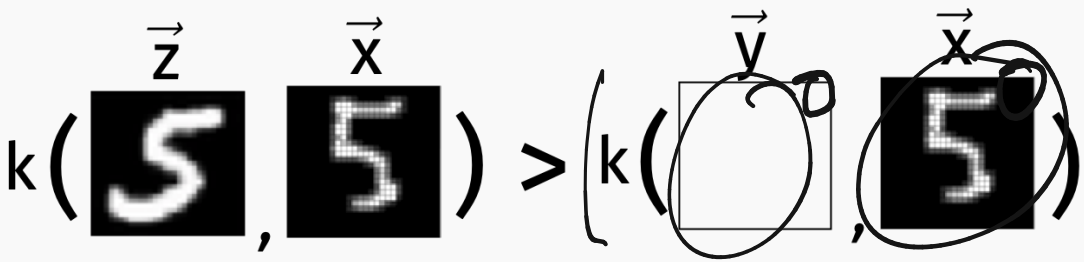
A kernel function $k(\mathbf{x}, \mathbf{y})$ is simply a similarity measure between data points.

$$k(\mathbf{x}, \mathbf{y}) = \begin{cases} \text{large if } \mathbf{x} \text{ and } \mathbf{y} \text{ are similar.} \\ \text{close to 0 if } \mathbf{x} \text{ and } \mathbf{y} \text{ are different.} \end{cases}$$

Example: The Radial Basis Function (RBF) kernel, aka the Gaussian kernel:

$$k(\mathbf{x}, \mathbf{y}) = e^{-\|\mathbf{x}-\mathbf{y}\|_2^2/\sigma^2}$$

for some scaling factor σ.



KERNEL FUNCTIONS: A NEW MEASURE OF SIMILARITY

Lots of kernel functions functions involve transformations of $\langle \mathbf{x}, \mathbf{y} \rangle$ or $\|\mathbf{x} - \mathbf{y}\|_2$:

- Gaussian RBF Kernel: $k(\mathbf{x}, \mathbf{y}) = e^{-\|\mathbf{x}-\mathbf{y}\|_2^2/\sigma^2}$
- Laplace Kernel: $k(\mathbf{x}, \mathbf{y}) = e^{-\|\mathbf{x}-\mathbf{y}\|_2/\sigma}$
- Polynomial Kernel: $k(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle + 1)^q$.

But you can imagine much more complex similarity metrics.

We will see one on the next problem set tailored to digit/letter recognition.

How to use a kernel function?

For k-nearest neighbors, can easily replace inner product with whatever similarity function you want.

For logistic regression, it is less clear how to do so.

How to use a kernel function?

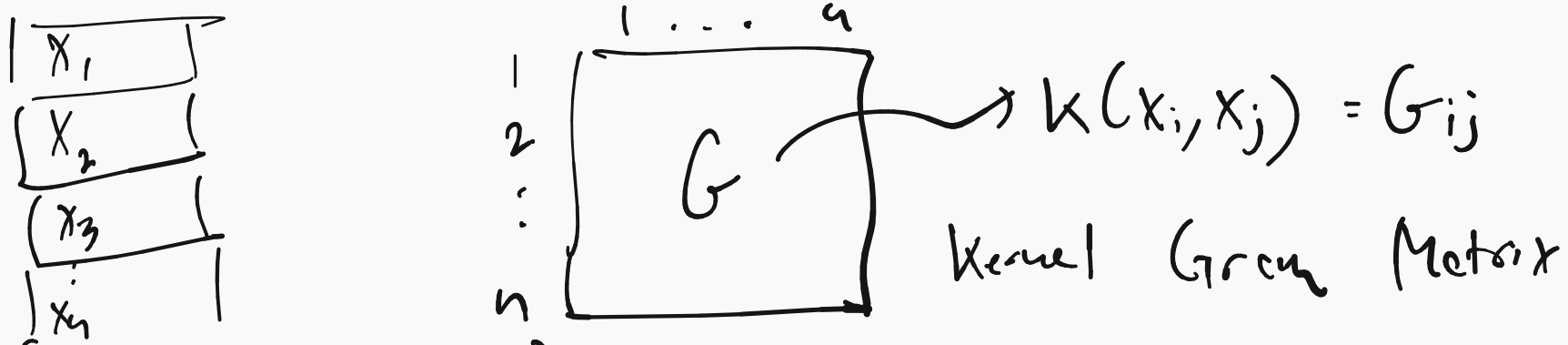
Logistic Regression Loss:

$$L(\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(q)}) = -\left(\sum_{i=1}^{n}\left(\sum_{\ell=1}^{q} \mathbb{1}[y_i = \ell]\right) \log \frac{e^{\langle \boldsymbol{\beta}^{(\ell)}, \mathbf{x}_i \rangle}}{\sum_{j=1}^{q} e^{\langle \boldsymbol{\beta}^{(j)}, \mathbf{x}_i \rangle}}\right)$$

Loss inherently involves inner product between each $\boldsymbol{\beta}^{(j)}$ and each data vector $\mathbf{x}_i$.

Solution: Only work with similarity metrics that can be expressed as inner products.

KERNEL FUNCTIONS FROM FEATURE TRANSFORMATION



A positive semidefinite (PSD) kernel is any similarity function with the following form:

$$k(\mathbf{x}, \mathbf{w}) = \phi(\mathbf{x})^T \phi(\mathbf{w}) = \langle \phi(x), \phi(w) \rangle$$

where $\phi : \mathbb{R}^d \to \mathbb{R}^m$ is a some feature transformation function.

$\phi(\quad) \to$

Example: Degree 2 polynomial kernel, $k(\mathbf{x}, \mathbf{w}) = (\mathbf{x}^T\mathbf{w} + 1)^2$.

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \qquad \phi(\mathbf{x}) = \begin{bmatrix} 1 \\ \sqrt{2}x_1 \\ \sqrt{2}x_2 \\ \sqrt{2}x_3 \\ x_1^2 \\ x_2^2 \\ x_3^2 \\ \sqrt{2}x_1x_2 \\ \sqrt{2}x_1x_3 \\ \sqrt{2}x_2x_3 \\ 0 \end{bmatrix}$$

$$\begin{aligned} (\mathbf{x}^T\mathbf{w} + 1)^2 &= (x_1w_1 + x_2w_2 + x_3w_3 + 1)^2 \\ &= 1 + 2x_1w_1 + 2x_2w_2 + 2x_3w_3 + x_1^2w_1^2 + x_2^2w_2^2 + x_3^2w_3^2 \\ &\quad + 2x_1w_1x_2w_2 + 2x_1w_1x_3w_3 + 2x_2w_2x_3w_3 \\ &= \phi(\mathbf{x})^T\phi(\mathbf{w}). \end{aligned}$$

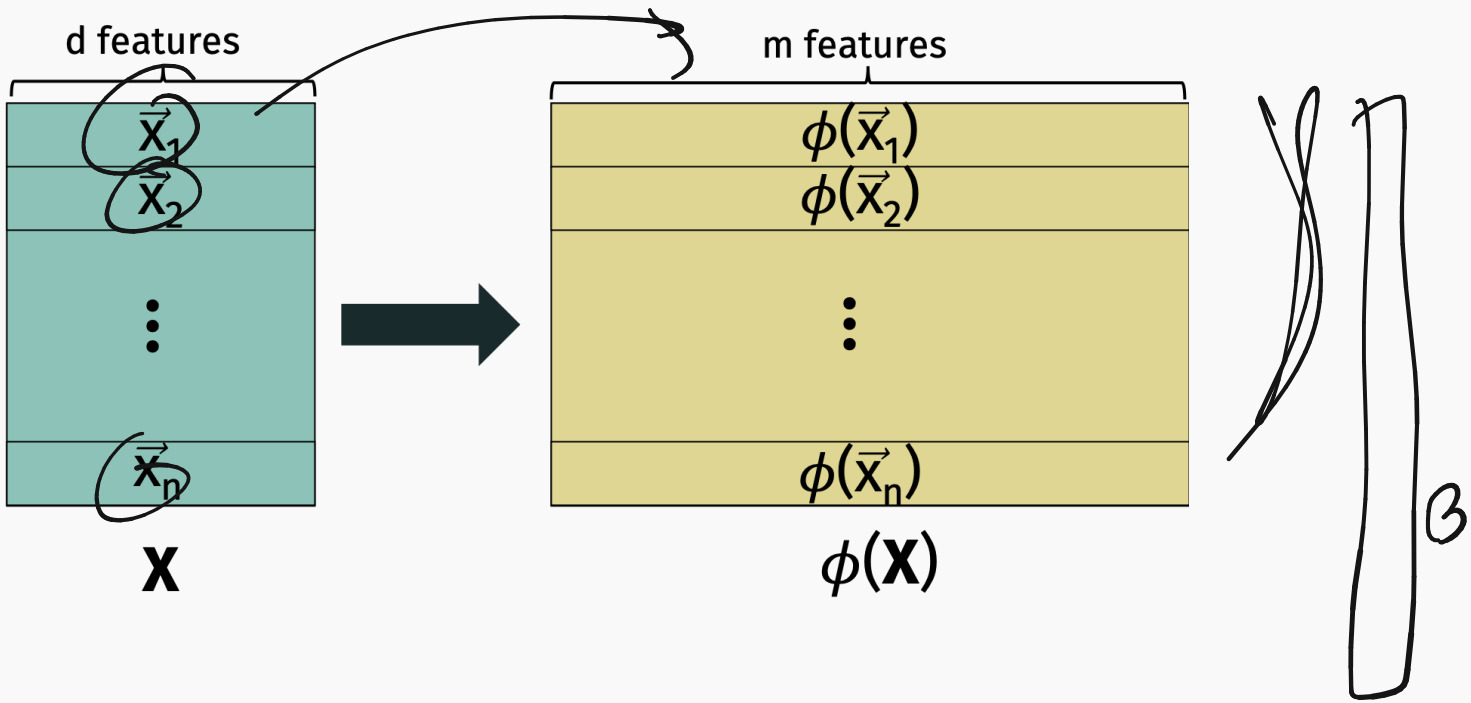
Not all similarity metrics are positive semidefinite (PSD), but all of the ones we saw earlier are:

- Gaussian RBF Kernel: $k(\mathbf{x}, \mathbf{y}) = e^{-\|\mathbf{x}-\mathbf{y}\|_2^2/\sigma^2}$
- Laplace Kernel: $k(\mathbf{x}, \mathbf{y}) = e^{-\|\mathbf{x}-\mathbf{y}\|_2/\sigma}$
- Polynomial Kernel: $k(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle + 1)^q$.

And there are many more...

Feature transformations $\Longleftrightarrow$ new similarity metrics.

To work with the similarity $k(\cdot, \cdot)$ in place of the inner product $\langle \cdot, \cdot \rangle$, it suffices to replace every data point $\mathbf{x}_1, \ldots, \mathbf{x}_n$ by $\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_n)$.



There are two major issues with this:

- While $\phi(\mathbf{x})$ is sometimes simple and explicit. **More often, it is not.** We might be able to show a kernel is PSD without easily being able to write down $\phi(\mathbf{x})$.
- Transform dimension m is often very large: e.g. $m = O(d^q)$ for a degree q polynomial kernel. For many kernels (e.g. the Gaussian kernel) m is actually *infinite*.

So doing the feature transformation explicitly would have very high computational cost.

REPARAMETERIZATION TRICK

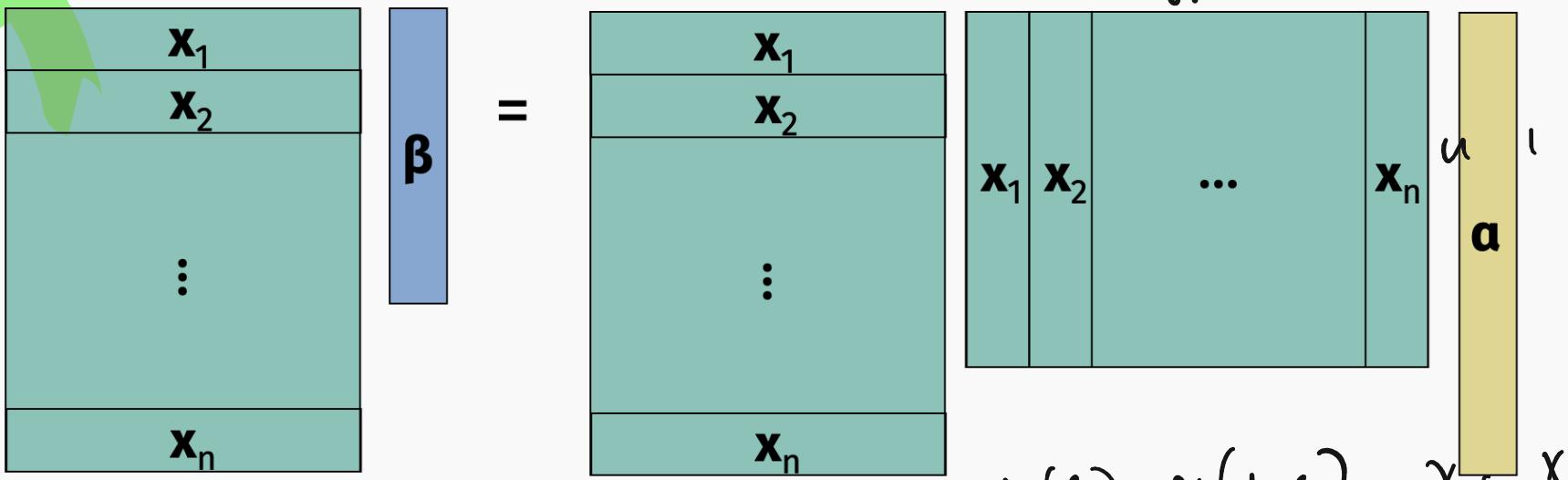
For simplicity, let's just consider the binary cross entropy/logistic regression loss:

$$-\sum_{j=1}^{n} y_j \log(h(\mathbf{X}\beta)_j) + (1 - y_j) \log(1 - h(\mathbf{X}\beta)_j)$$

where $h(z) = \frac{1}{1+e^{-z}}$.

Reminder from linear algebra: Without loss of generality, can assume that β lies in the row span of X.

So for any $\beta \in \mathbb{R}^d$, there exists a vector $\alpha \in \mathbb{R}^n$ such that: $X\beta = XX^T\alpha$.



REPARAMETERIZATION TRICK

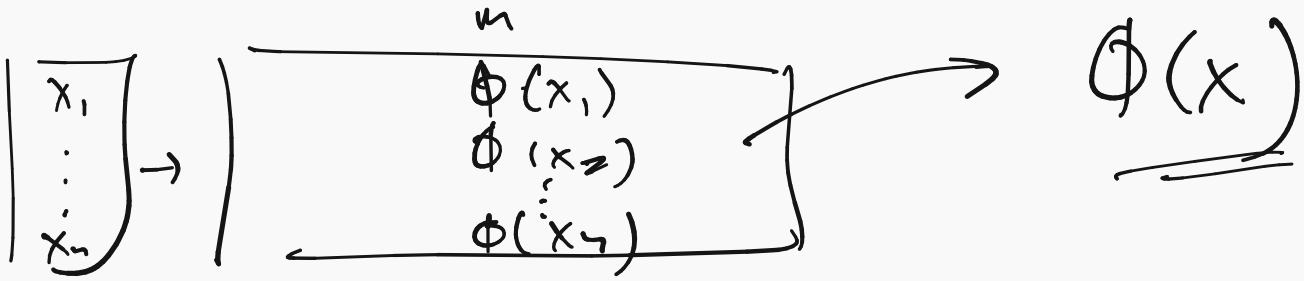
Logistic Regression Equivalent Formulation: Given data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and binary label vector $\mathbf{y} \in \{0,1\}^n$ for class i, find $\boldsymbol{\alpha} \in \mathbb{R}^n$ to minimize the loss:

$$-\sum_{j=1}^{n} y_j \log(h(\mathbf{X}\mathbf{X}^T\boldsymbol{\alpha})_j) + (1-y_j)\log(1-h(\mathbf{X}\mathbf{X}^T\boldsymbol{\alpha})_j)$$

Can still be minimized via gradient descent:

$$\nabla L(\boldsymbol{\alpha}) = \mathbf{X}\mathbf{X}^T(h(\mathbf{X}\mathbf{X}^T\boldsymbol{\alpha}) - \mathbf{y}).$$

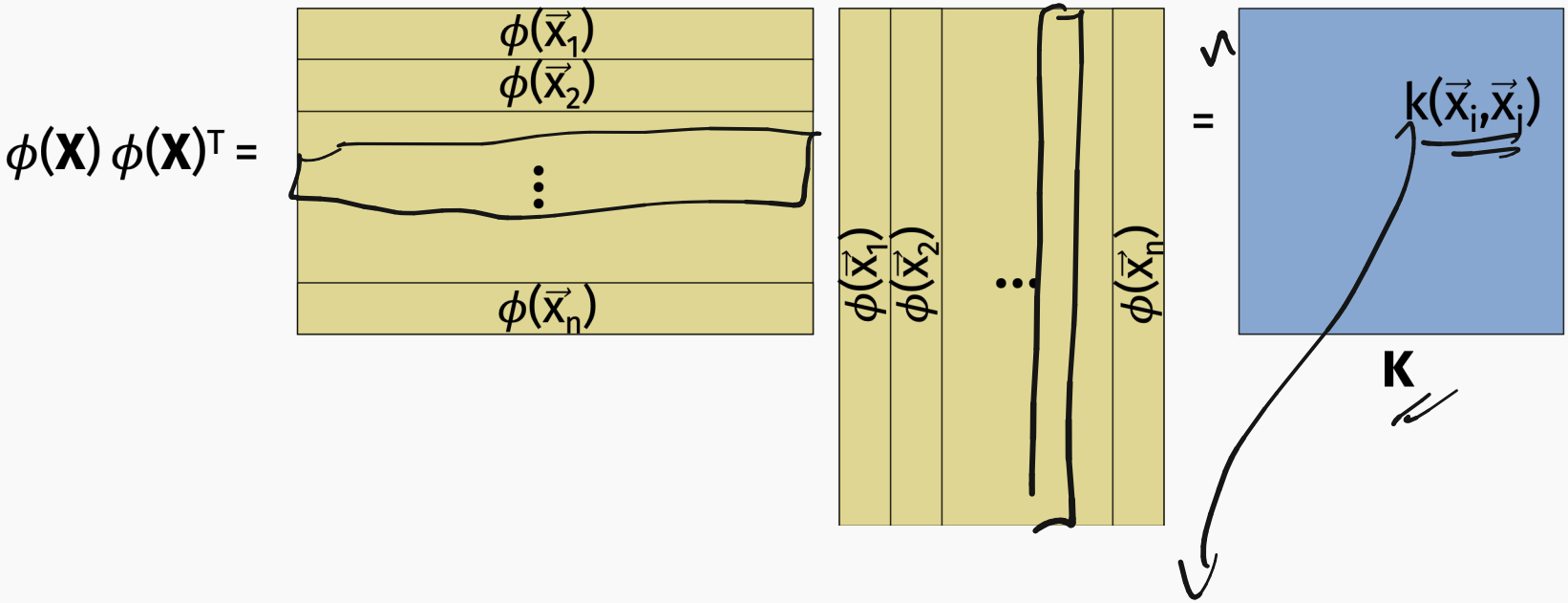
REPARAMETERIZATION TRICK



If we use a non-linear data transformation ϕ (corresponding to a PSD kernel), then the loss is:

$$-\sum_{j=1}^{n} y_j \log(h(\phi(\mathbf{X})\phi(\mathbf{X})^T\boldsymbol{\alpha})_j) + (1-y_j)\log(1-h(\phi(\mathbf{X})\phi(\mathbf{X})^T\boldsymbol{\alpha})_j)$$

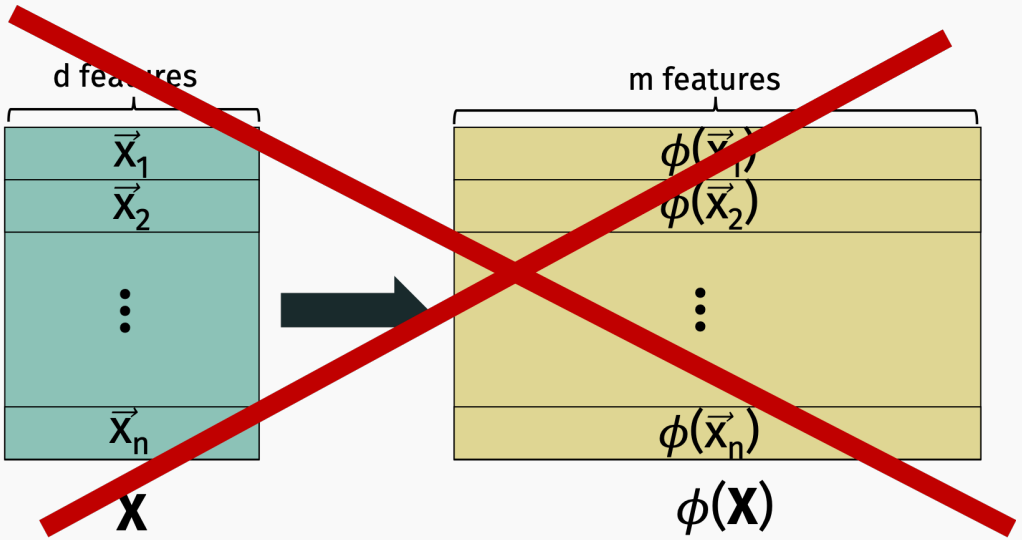
$\mathbf{K} = \phi(\mathbf{X})\phi(\mathbf{X})^T$ is called the kernel Gram matrix.



$\langle \phi(x_i), \phi(x_j) \rangle = k(x_i, x_j)$

We never need to actually compute $\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_n)$ explicitly!

- For training we just need the kernel matrix $\mathbf{K}$, which requires computing $k(\mathbf{x}_i, \mathbf{x}_j)$ for all i, j.



We can always work with a finite sized $n \times n$ matrix.

Take away:

- Logistic regression can be combined with any positive semidefinite kernel matrix, and the model can be trained in time independent of the transform dimension m.

Prediction can also be done efficiently. For a new input $\mathbf{x}_{new}$, we need to compute:

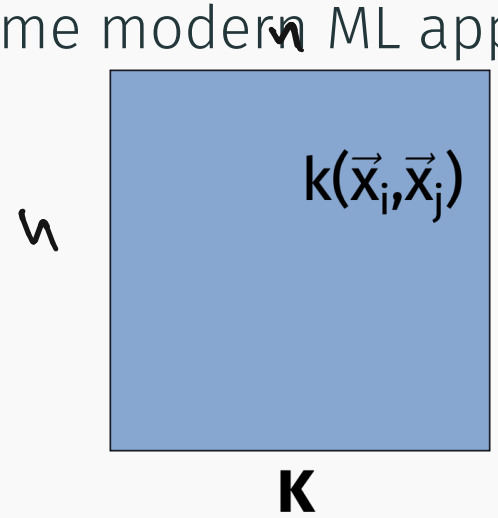
$$\langle \phi(\mathbf{x}_{new}), \beta \rangle = \langle \phi(\mathbf{x}_{new}), \phi(\mathbf{X})^\top \alpha \rangle = \sum_{j=1}^{n} \alpha_j \langle \phi(\mathbf{x}_{new}), \phi(\mathbf{x}_j) \rangle.$$

Each term in the sum $\langle \phi(\mathbf{x}_{new}), \phi(\mathbf{x}_j) \rangle = k(\mathbf{x}_{new}, \mathbf{x}_j)$ can be computed without explicit feature transformation.[2]

[2] Note that this does require computing the kernel inner product with potentially all examples in the training data. More on this shortly.

BEYOND THE KERNEL TRICK

The kernel matrix $\mathbf{K}$ is still $n \times n$ though which is huge when the size of the training set n is large. Has made the kernel trick less appealing in some modern ML applications.

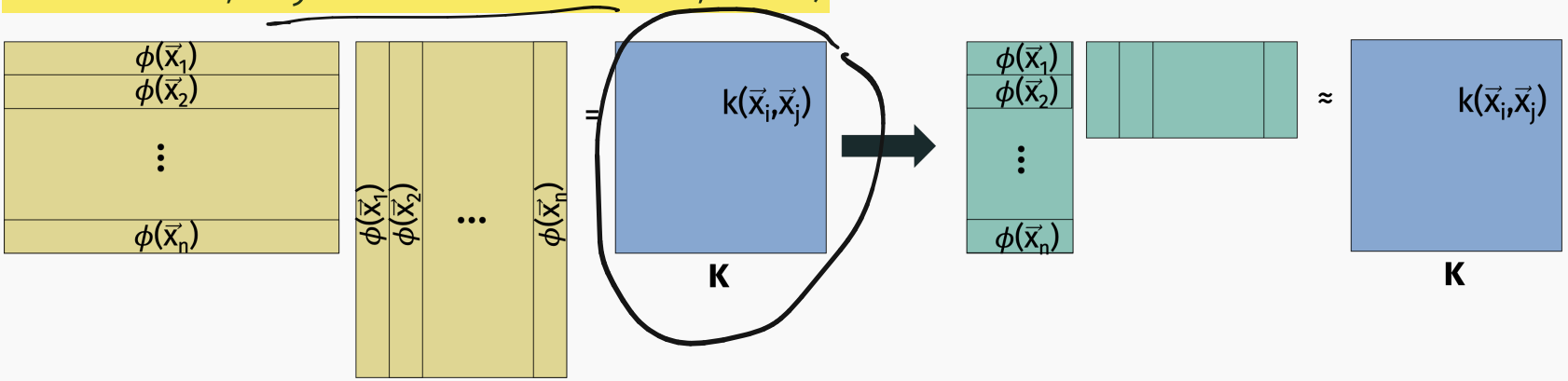


There is an inherent quadratic dependence on n in the computational and space complexity of kernel methods.

- $10,000$ data points $\rightarrow$ runtime scales as $\sim 100,000,000$, $\mathbf{K}$ takes 800MB of space.
- $1,000,000$ data points $\rightarrow$ runtime scales as $\sim 10^{12}$, $\mathbf{K}$ takes 8TB of space.

BEYOND THE KERNEL TRICK

Many algorithmic advances in recent years partially address this computational challenge (random Fourier features methods, Nystrom methods, etc.)



Often based on “reversing” the kernel trick to find a compact feature set that well approximates the kernel.

The kernel trick can also be applied outside of classification.

E.g. to regression:

$$\min_{\boldsymbol{\beta}} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2 \rightarrow \min_{\boldsymbol{\alpha}} \|\mathbf{X}\mathbf{X}^T\boldsymbol{\alpha} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{X}^T\boldsymbol{\alpha}\|_2^2$$

Replace $\mathbf{X}\mathbf{X}^T$ by kernel matrix $\mathbf{K}$ during training.

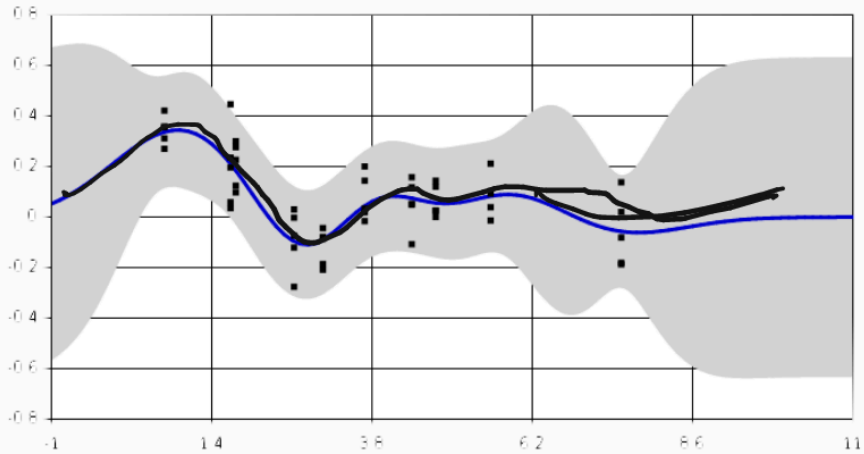
Prediction:

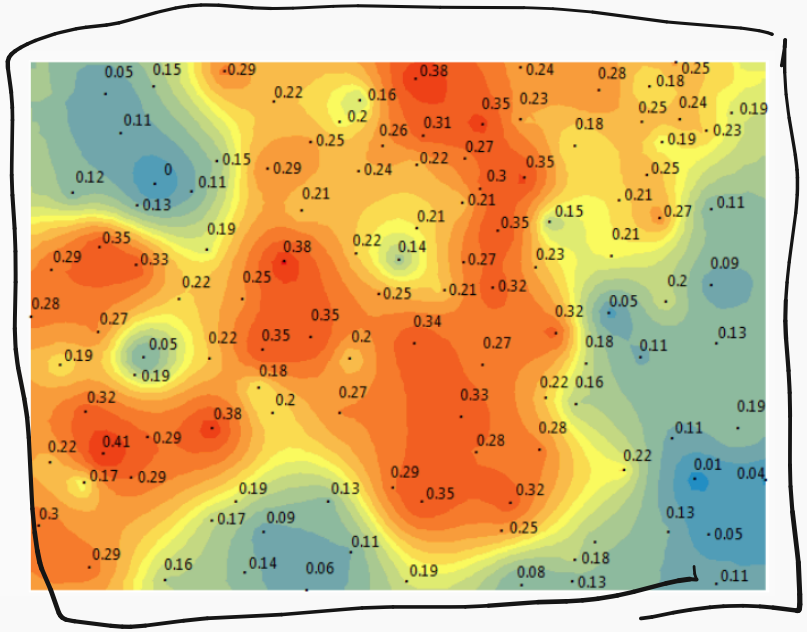
$$y_{new} = \sum_{i=1}^{n} \alpha_i \cdot k(\mathbf{x}_{new}, \mathbf{x}_i).$$

Added benefit: Relatively numerically stable. E.g. is a much better option for performing multivariate or even single variate polynomial regression than direct feature expansion.

KERNEL REGRESSION

Kernel regression with non-linear kernels like $e^{-\|\mathbf{x}-\mathbf{y}\|_2^2}$ is a very important statistical tool, especially when dealing with spatial or temporal data.





Also known as Gaussian Process (GP) Regression or Kriging. Most commonly, Gaussian kernel $k(\mathbf{x}, \mathbf{y}) = e^{-\|\mathbf{x}-\mathbf{y}\|_2^2/\sigma^2}$ is used in place of, e.g., polynomial kernel/polynomial regression.

SUPPORT VECTOR MACHINES

TODAY

Support Vector Machines (SVMs): Another algorithm for finding linear classifiers which is (was?) as popular as logistic regression.

- Can also be combined with kernels.
- Developed from a pretty different perspective.
- But final algorithm is not that different.

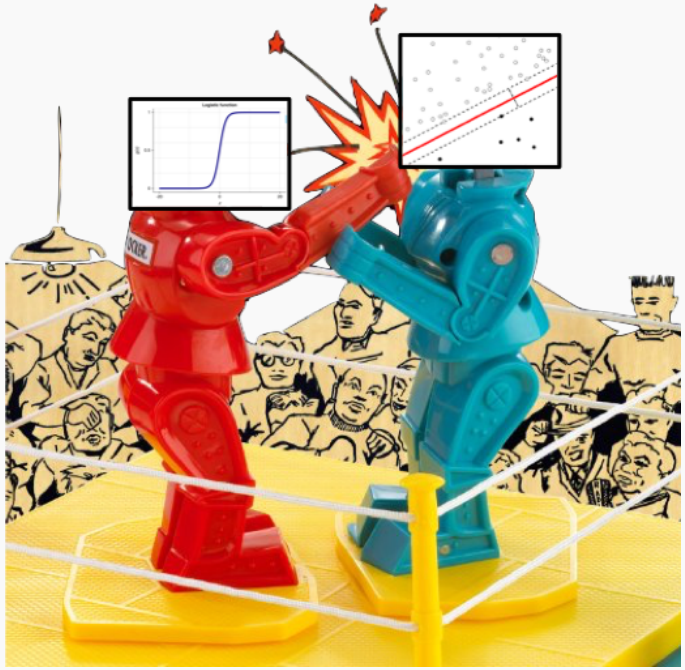
- Invented in 1963 by Alexey Chervonenkis and Vladimir Vapnik. Also founders of VC-theory.
- First combined with non-linear kernels in 1993.

SVM's vs. logistic regression

SVMs are more commonly associated with non-linear kernels. For example, `sklearn`'s SVM classifier (called SVC) has support for non-linear kernels built in by default. Its logistic regression classifier does not.

- Seeems to be partially for historical reasons.
- In the early 2000s SVMs where a "hot topic" in machine learning and their popularity persists.
- It is not clear to me if they are better than logistic regression, but honestly the jury is still out...
- There are some computational advantages of using SVMs, and some disadvantages, which is part of the story.

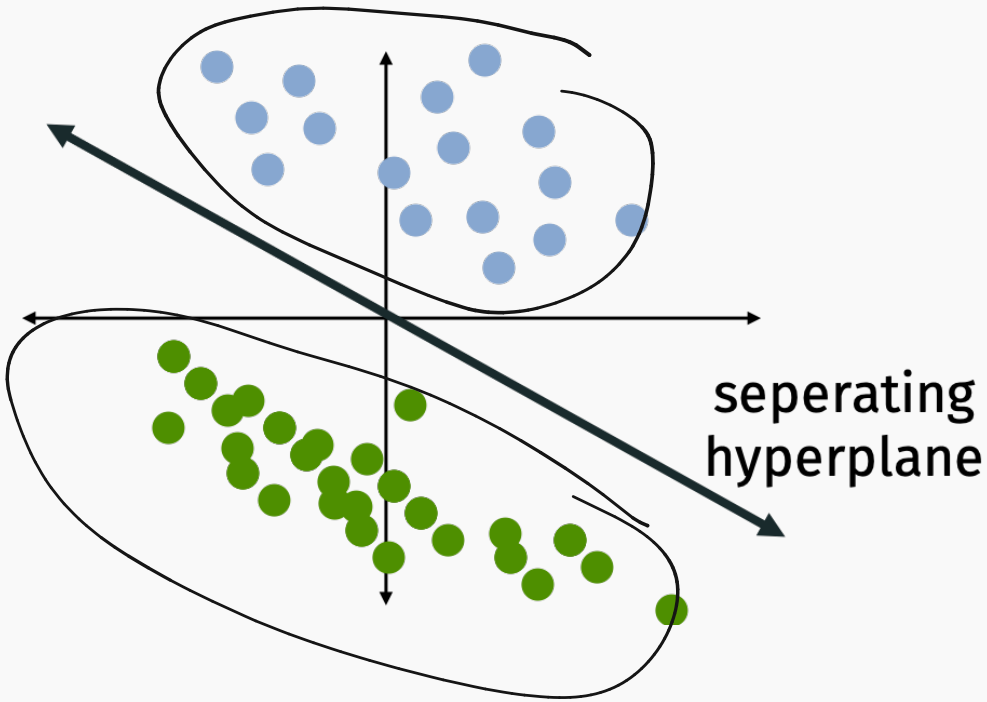
SVM's vs. logistic regression



Next lab: Machina-a-machina comparison of SVMs vs. logistic regression for a MNIST digit classification problem. Which provides better accuracy? Which is faster to train?

LINEARLY SEPARABLE DATA

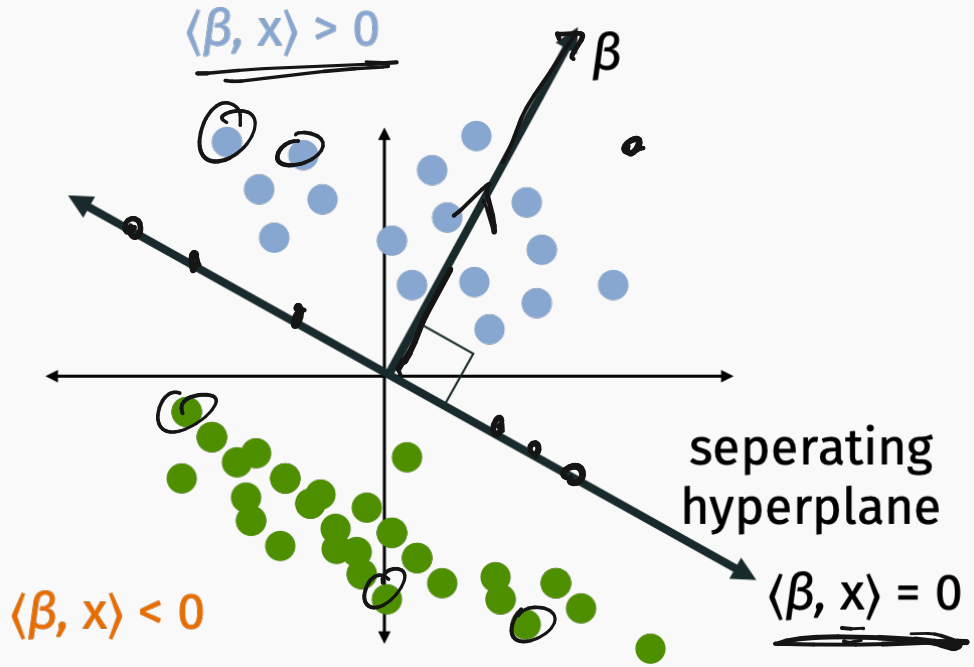
We call a dataset with binary labels linearly separable if it can be perfectly classified with a linear classifier:



This the realizable setting we discussed last lecture.

LINEARLY SEPARABLE DATA

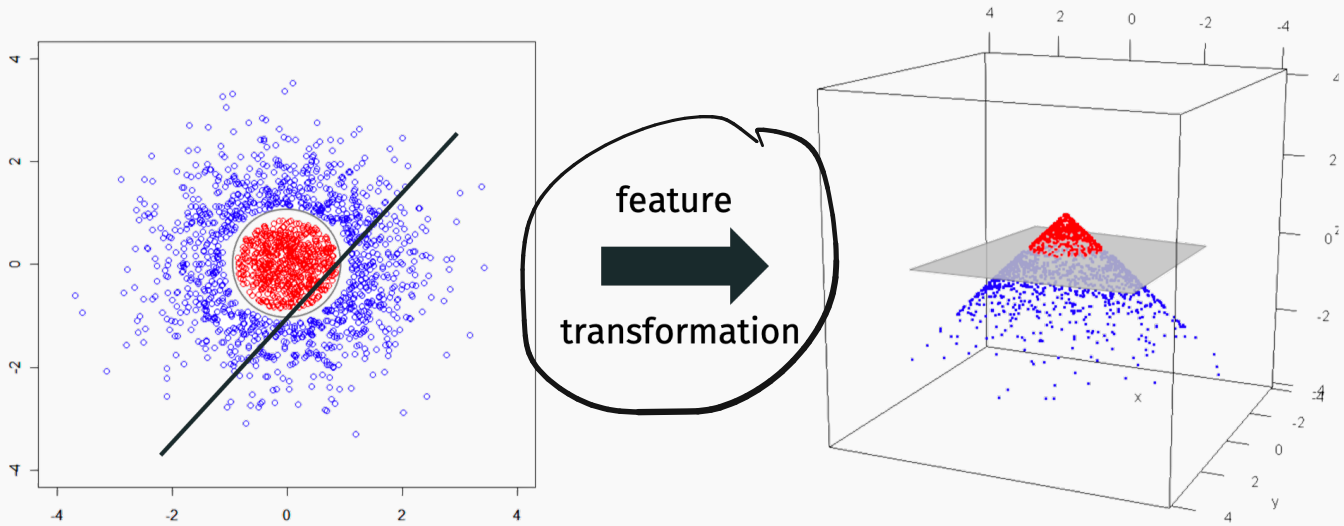
Formally, there exists a parameter β such that $\langle \beta, \mathbf{x} \rangle > 0$ for all $\mathbf{x}$ in class 1 and $\langle \beta, \mathbf{x} \rangle < 0$ for all $\mathbf{x}$ in class 0.



Note that if we multiply β by any constant c, $c\beta$ gives the same separating hyperplane because $\langle c\beta, \mathbf{x} \rangle = c\langle \beta, \mathbf{x} \rangle$.

LINEARLY SEPARABLE DATA

A data set might be linearly separable when using a non-kernel/feature transformation even if it is not separable in the original space.

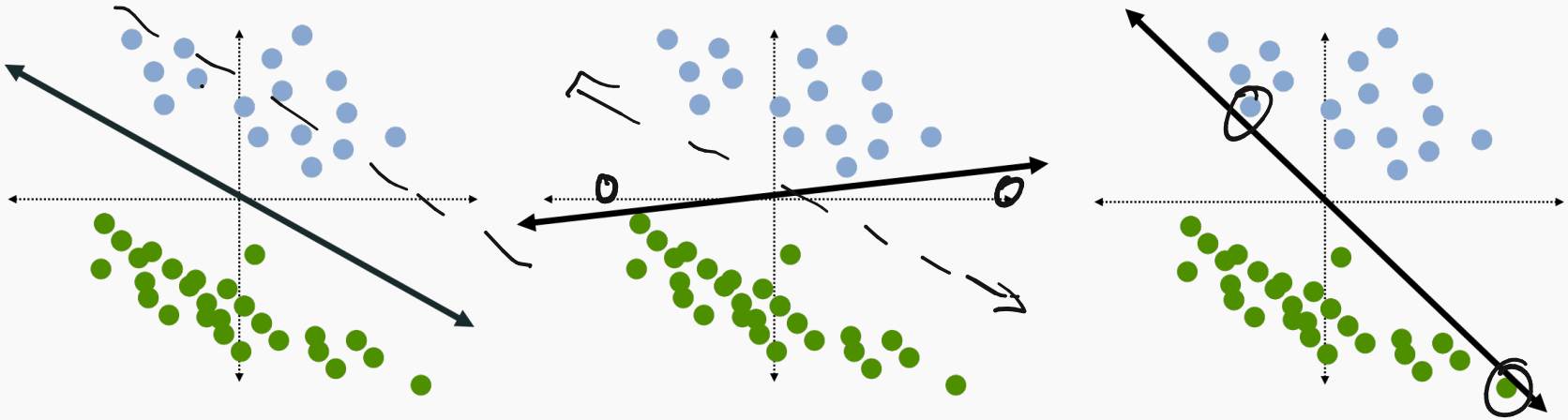


This data is separable when using a degree-2 polynomial kernel. If suffices for $\phi(\mathbf{x})$ to contain x_1^2 and x_2^2.

MARGIN

When data is linearly separable, we would typically be concerned about over-fitting.

Idea from Vapnik and Chervonenkis: Maybe there is a way to find a classification rule that over-fits on the training data but is still "good" for future data.

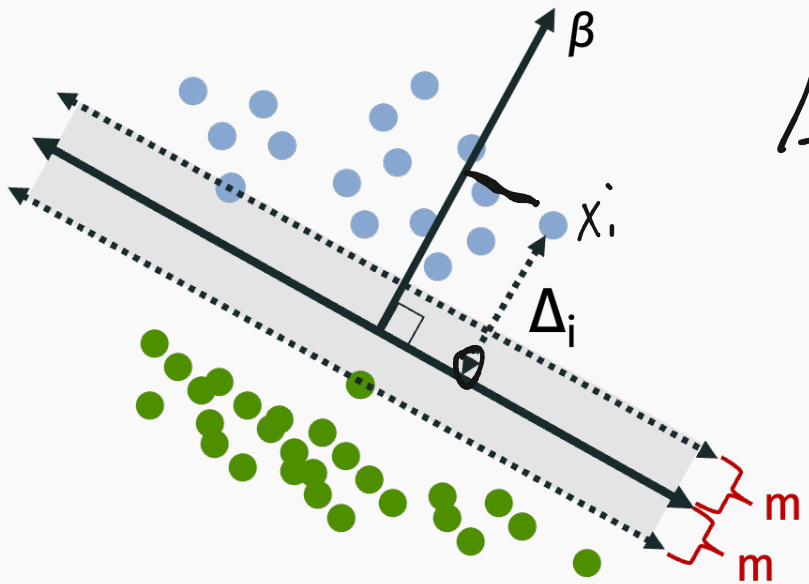


There are typically multiple valid separating hyperplanes.

Intuitively, which is best for classifying future data?

MARGIN

The **margin** m of a separating hyperplane is the minimum ℓ_2 (Euclidean) distance between a point in the dataset and the hyperplane.



L2 distance

$$m = \min_i \Delta_i \qquad \text{where} \qquad \Delta_i = \frac{|\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle|}{\|\boldsymbol{\beta}\|_2}$$

We have that $\mathbf{x}_i = \mathbf{v}_i + \mathbf{e}_i$ where $\mathbf{v}_i$ is parallel to $\boldsymbol{\beta}$ and $\mathbf{e}_i$ is perpendicular.

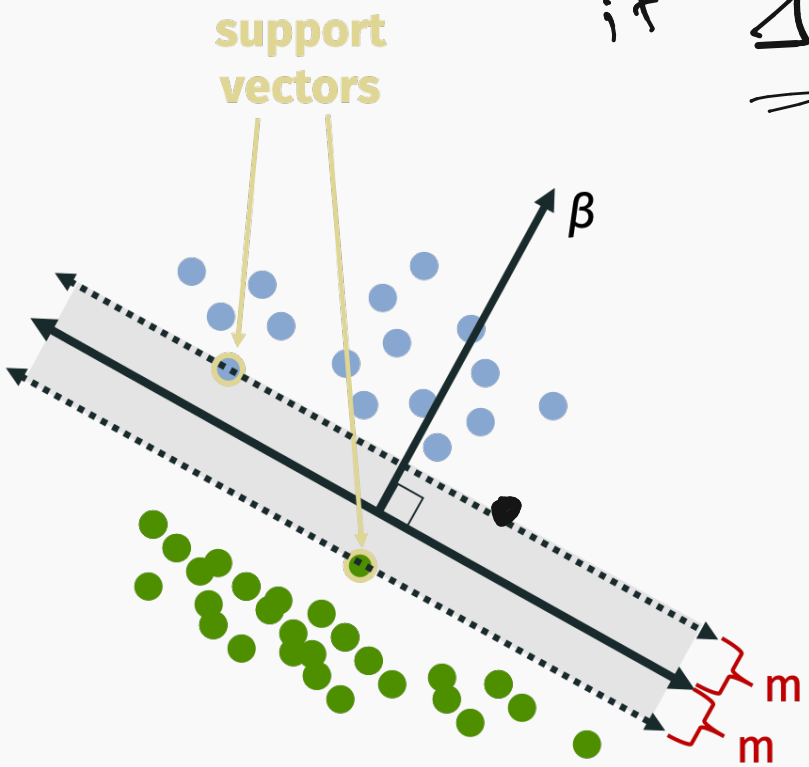
$$\Delta_i = \|\mathbf{v}_i\|_2 = \frac{1}{\|\mathbf{v}_i\|_2} \cdot \langle \mathbf{v}_i, \mathbf{v}_i \rangle = \frac{1}{\|\mathbf{v}_i\|_2} \cdot \frac{\|\mathbf{v}_i\|_2}{\|\boldsymbol{\beta}\|_2} \cdot |\langle \mathbf{v}_i, \boldsymbol{\beta}_i \rangle| = \frac{|\langle \mathbf{v}_i, \boldsymbol{\beta} \rangle|}{\|\boldsymbol{\beta}\|_2}.$$

Finally, we have that $\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle = \langle \mathbf{v}_i, \boldsymbol{\beta} \rangle$ because $\langle \mathbf{e}_i, \boldsymbol{\beta} \rangle = 0$.

SUPPORT VECTOR

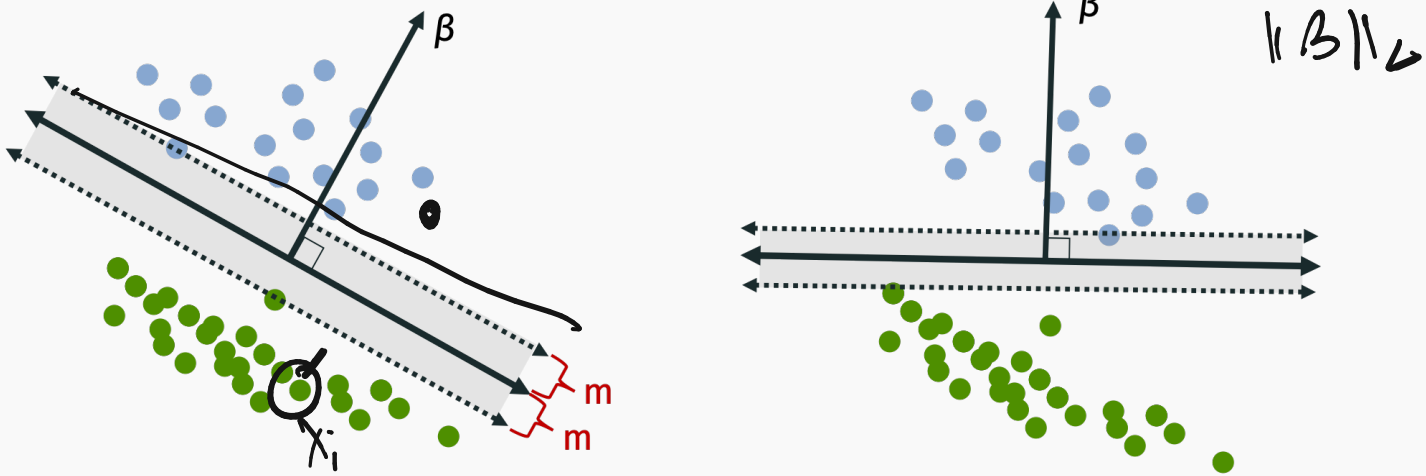
A **support vector** is any data point $\mathbf{x}_i$ such that $\frac{|\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle|}{\|\boldsymbol{\beta}\|_2} = m$, where $m = \min_i \frac{|\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle|}{\|\boldsymbol{\beta}\|_2}$.

x_i is support vector if $\Delta_i = m$



HARD-MARGIN SVM

A (hard-margin) support vector machine (SVM) classifier finds the **maximum margin (MM) linear classifier**.



I.e. the separating hyperplane which maximizes the margin m. Like regularization, doing so can prevent overfitting even if we can fit our data perfectly.

Denote the maximum margin by m^*.

$$m^* = \max_{\beta} \left[\min_{i \in 1,\dots,n} \frac{|\langle \mathbf{x}_i, \beta \rangle|}{\|\beta\|_2} \right]$$
$$= \max_{\beta} \left[\min_{i \in 1,\dots,n} \frac{y_i \cdot \langle \mathbf{x}_i, \beta \rangle}{\|\beta\|_2} \right]$$

where $y_i = -1, 1$ depending on what class $\mathbf{x}_i$ is in.[3]

[3] Note that this is a different convention than the $0, 1$ class labels we typically use.

HARD-MARGIN SVM

Original problem: $\max_\beta \left[\min_{i \in 1,\dots,n} \frac{y_i \cdot \langle \mathbf{x}_i, \beta \rangle}{\|\beta\|_2} \right]$.

Equivalent formulation:

$$\min_\beta \|\beta\|_2^2 \quad \text{subject to} \quad y_i \cdot \langle \mathbf{x}_i, \beta \rangle \geq 1 \text{ for all } i.$$

$$\max_\beta \left[\min_i \frac{y_i \cdot \langle x_i, \beta \rangle}{\|\beta\|_2} \right] = \frac{1}{\|\beta\|_2}$$

$y_i \cdot \langle x_i, \beta \rangle \geq 1$ for all i

and $y_i \cdot \langle x_i, \beta \rangle = 1$ for some i.

$$\max_\beta \frac{1}{\|\beta\|_2}$$

Training data:

$\langle x_i, \beta \rangle \geq 1$ for all i s.t. $y_i = 1$

$\langle x_i, \beta \rangle \leq -1$ for all i s.t. $y_i = -1$

$\mathbb{1}[\langle x_{new}, \beta \rangle > 0]$

Equivalent formulation:

$$\min_{\beta} \|\beta\|_2 \quad \text{subject to} \quad y_i \cdot \langle \mathbf{x}_i, \beta \rangle \geq 1 \text{ for all } i.$$

Under this formulation $m^* = \frac{1}{\|\beta\|_2}$.

This is a **constrained optimization problem.** In particular, a linearly constrained quadratic program, which is a type of problem we have efficient optimization algorithms for.

Not as easy to solve as a standard unconstrained, convex optimization problem like logistic regression!

Classification rule is the same as usual: classify in class 1 if $\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle \geq 0$, class -1 if $\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle$.

Kernel case: As before, we can parameterize as $\boldsymbol{\beta} = \mathbf{X}^T \boldsymbol{\alpha}$. When using a non-linear kernel $k(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^T \phi(\mathbf{y})$, we have:

$$\boldsymbol{\beta} = \phi(\mathbf{X})^T \boldsymbol{\alpha}$$

and to classify a new point $\mathbf{x}_{new}$ we compute:

$$\phi(\mathbf{x}_{new})^T \boldsymbol{\beta} = \phi(\mathbf{x}_{new})^T \phi(\mathbf{X})^T \boldsymbol{\alpha} = \phi(\mathbf{x}_{new})^T \sum_{i=1}^{n} \alpha_i \phi(\mathbf{x}_i) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_{new}, \mathbf{x}_i).$$

Can show that $\alpha_i = 0$ whenever $\mathbf{x}_i$ is not a support vector.

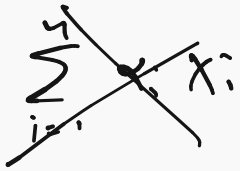
Classification cost scales with # of support vectors, s, not n.

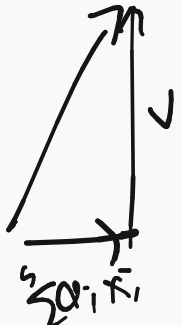
HARD-MARGIN SVM CLASSIFICATION

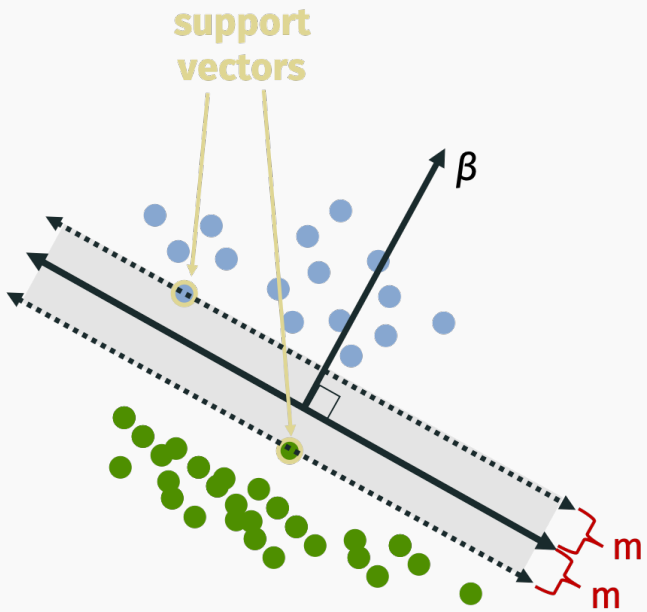
$$\min_{\beta} \|\beta\|_2 \quad \text{subject to} \quad y_i \cdot \langle \mathbf{x}_i, \beta \rangle \geq 1 \text{ for all } i.$$

Claim: Let $\mathbf{x}_1, \ldots, \mathbf{x}_s$ be all vectors with $y_i \cdot \langle \mathbf{x}_i, \beta \rangle = 1$. $s \ll n$

$\beta = \sum_{i=1}^{s} \alpha_i \mathbf{x}_i$.







i : $\mathbf{x}_i$ is a support vector

$y_i \langle \mathbf{x}_i, \beta \rangle = y_i \langle \mathbf{x}_i, \beta' \rangle$

$y_i \langle \mathbf{x}_i, \beta - \epsilon \mathbf{v} \rangle$

$= y_i \langle \mathbf{x}_i, \beta \rangle$

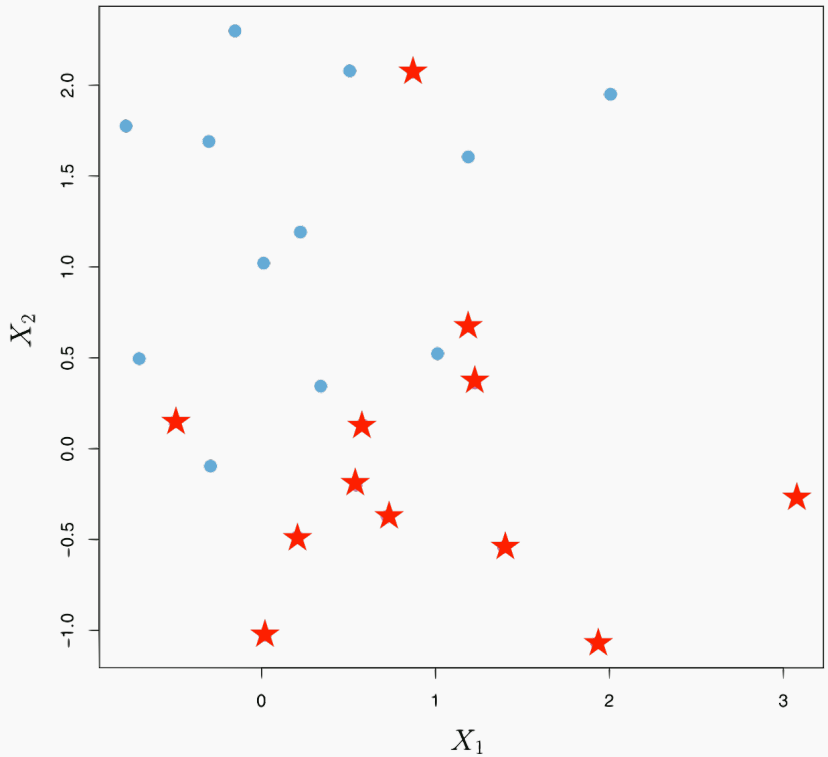
Proof by contradiction: Suppose $\beta = \sum_{i=1}^{s} \alpha_i \mathbf{x}_i + \mathbf{v}$, where $\mathbf{v}$ is orthogonal to the support vectors. Can reduced $\|\beta\|_2$ by setting $\beta' \leftarrow \sum_{i=1}^{s} \alpha_i \mathbf{x}_i + (1-\epsilon)\mathbf{v}$ for some small ϵ

HARD-MARGIN SVM CLASSIFICATION

Take-away:

- Training SVMs is typically harder than training using logistic loss.
- Classification after the model is trained requires $O(n)$ kernel evaluations for general linear classifiers (e.g., found via logistic regression), but just $O(s)$ for an SVM with s support vectors. Often, $s \ll n$.
- Advantages in-terms of storage space as well: only need to keep support vectors around for classification.

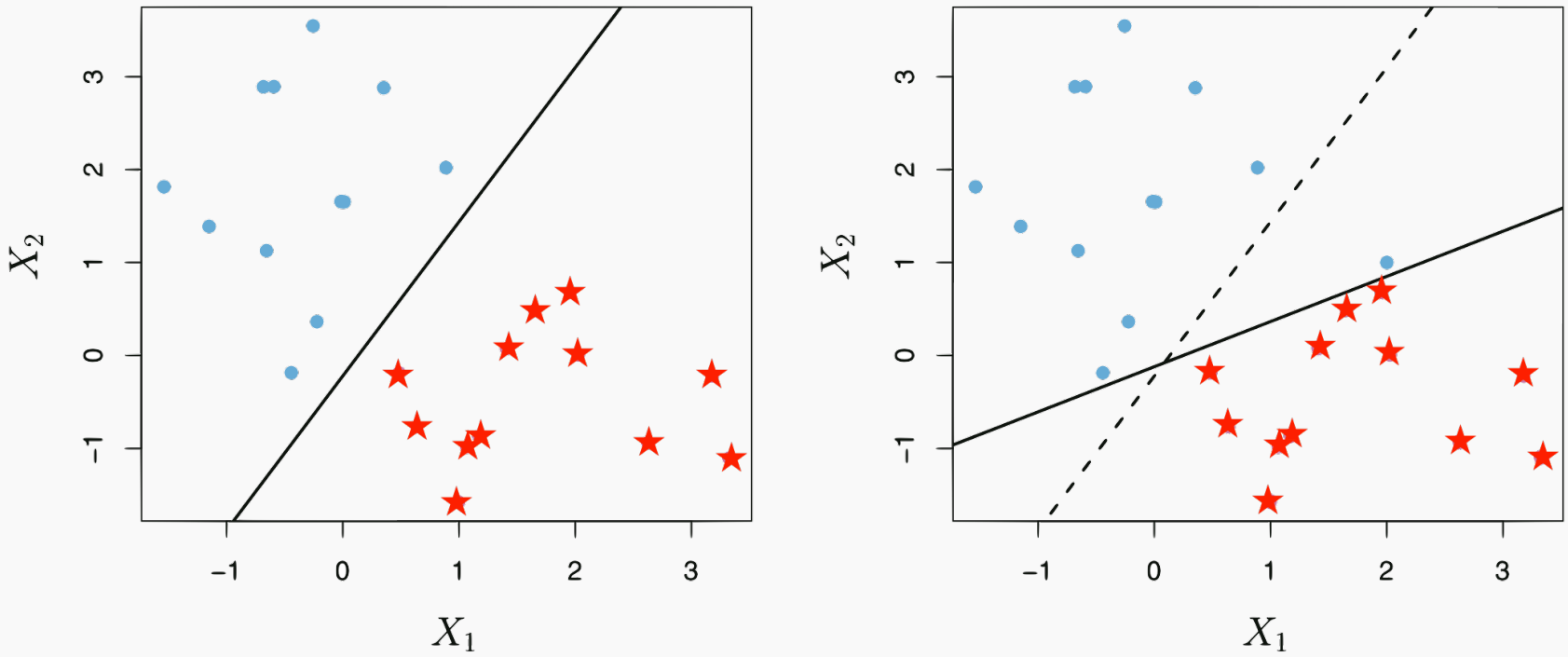
Hard-margin SVMs have a few other critical issues in practice:



Data might not be linearly separable, in-which case the maximum margin classifier is not even defined.

Less likely to be an issue when using a non-linear kernel. If $\mathbf{K}$ is full rank then perfect separation is always possible. And typically it is, e.g. for an RBF kernel or moderate degree polynomial kernel.

Another critical issue in practice:



Hard-margin SVM classifiers are not robust.

Solution: Allow the classifier to make some "mistakes"! A mistake can either be a misclassification, or simply a point allowed to be "inside" the margin.

Hard margin objective:

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_2^2 \qquad \text{subject to} \qquad y_i \cdot \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle \geq 1 \text{ for all } i.$$

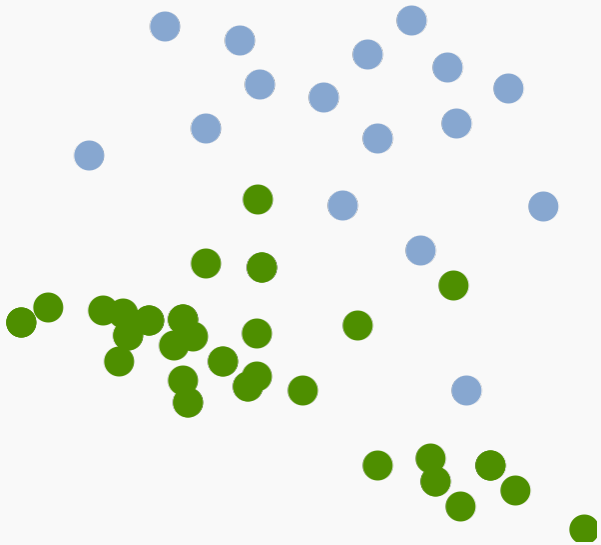
Soft margin objective:

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_2^2 + C \sum_{i=1}^{n} \epsilon_i \quad \text{subject to} \quad y_i \cdot \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle \geq 1 - \epsilon_i \text{ for all } i.$$

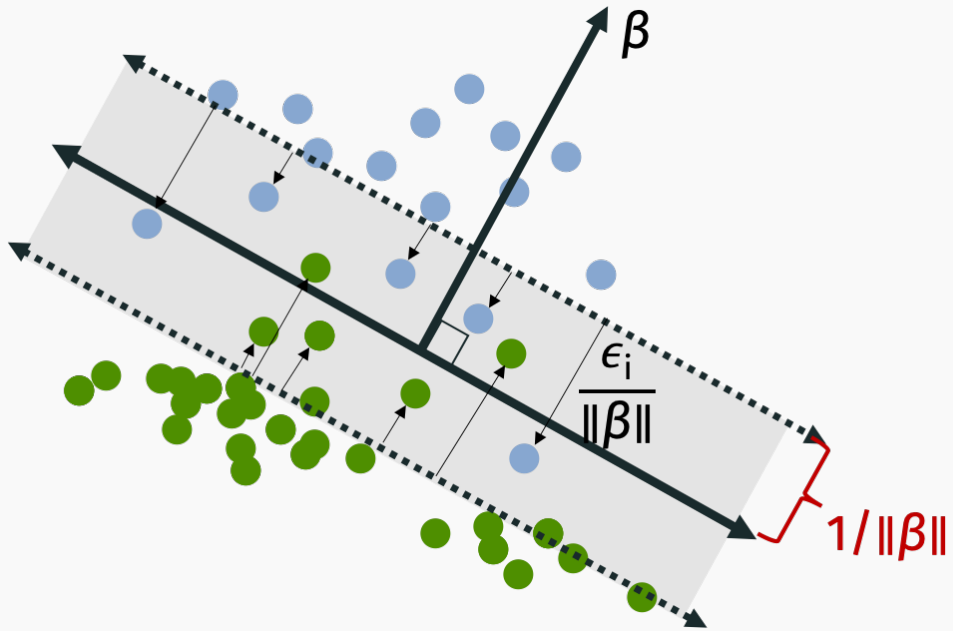
where $\epsilon_i \geq 0$ is a non-negative "slack variable". This is the magnitude of the "error" (distance past the margin) we allow example $\mathbf{x}_i$ to travel. $\epsilon_i \geq 1$ corresponds to a misclassification.

$C \geq 0$ is a non-negative tuning parameter.

Example of a non-separable problem:



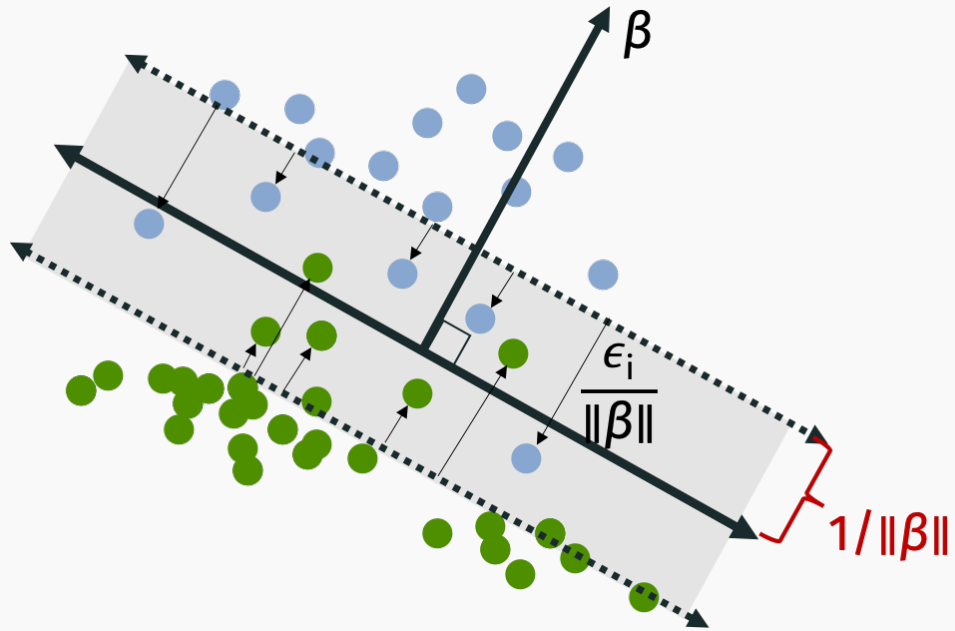
Recall that $\Delta_i = \frac{y_i \cdot \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle}{\|\boldsymbol{\beta}\|_2}$.



Soft margin objective:

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_2^2 + C \sum_{i=1}^{n} \epsilon_i \quad \text{subject to} \quad y_i \cdot \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle \geq 1 - \epsilon_i \text{ for all } i.$$

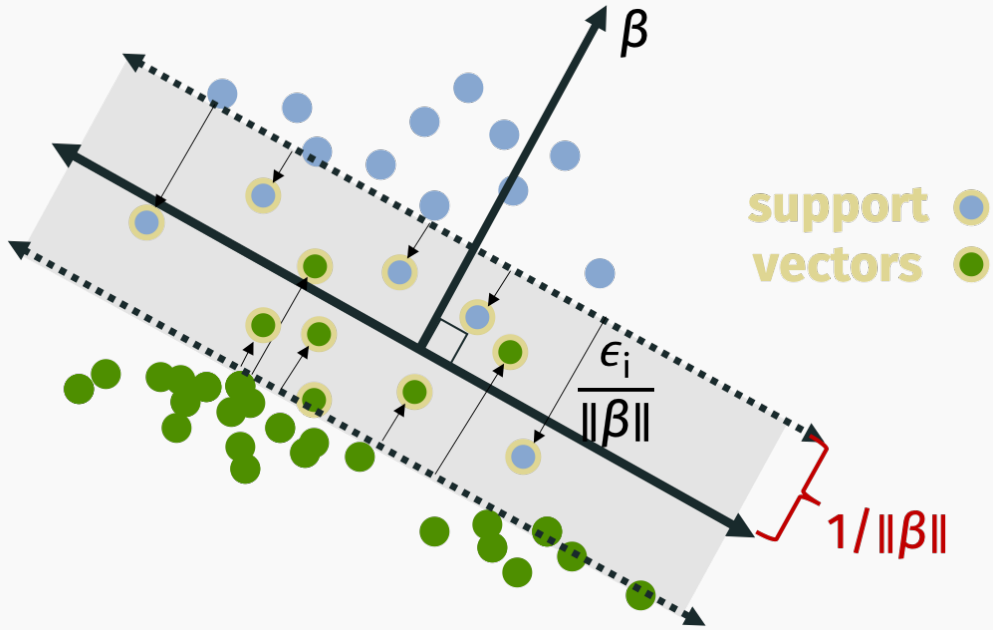
Recall that $\Delta_i = \frac{y_i \cdot \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle}{\|\boldsymbol{\beta}\|_2}$.



Soft margin objective:

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_2^2 + C \sum_{i=1}^{n} \epsilon_i \quad \text{subject to} \quad \frac{y_i \cdot \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle}{\|\boldsymbol{\beta}\|_2} \geq \frac{1}{\|\boldsymbol{\beta}\|_2} - \frac{\epsilon_i}{\|\boldsymbol{\beta}\|_2} \text{ for all } i.$$

SOFT-MARGIN SVM



Any $\mathbf{x}_i$ with a non-zero ϵ_i is a support vector. As before, only support vectors are needed for classification in the kernel setting.

Soft margin objective:

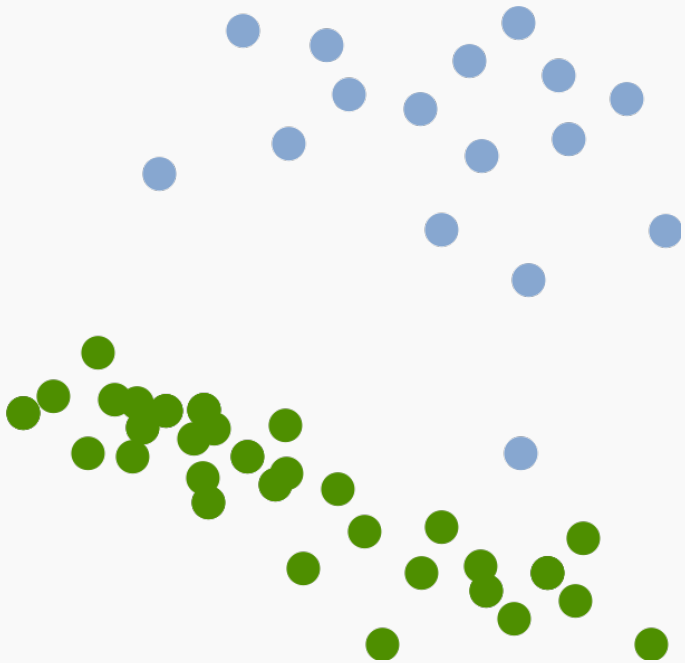
$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_2^2 + C \sum_{i=1}^{n} \epsilon_i.$$

- Large C means penalties are punished more in objective $\implies$ smaller margin, less support vectors.
- Small C means penalties are punished less in objective $\implies$ larger margin, more support vectors.

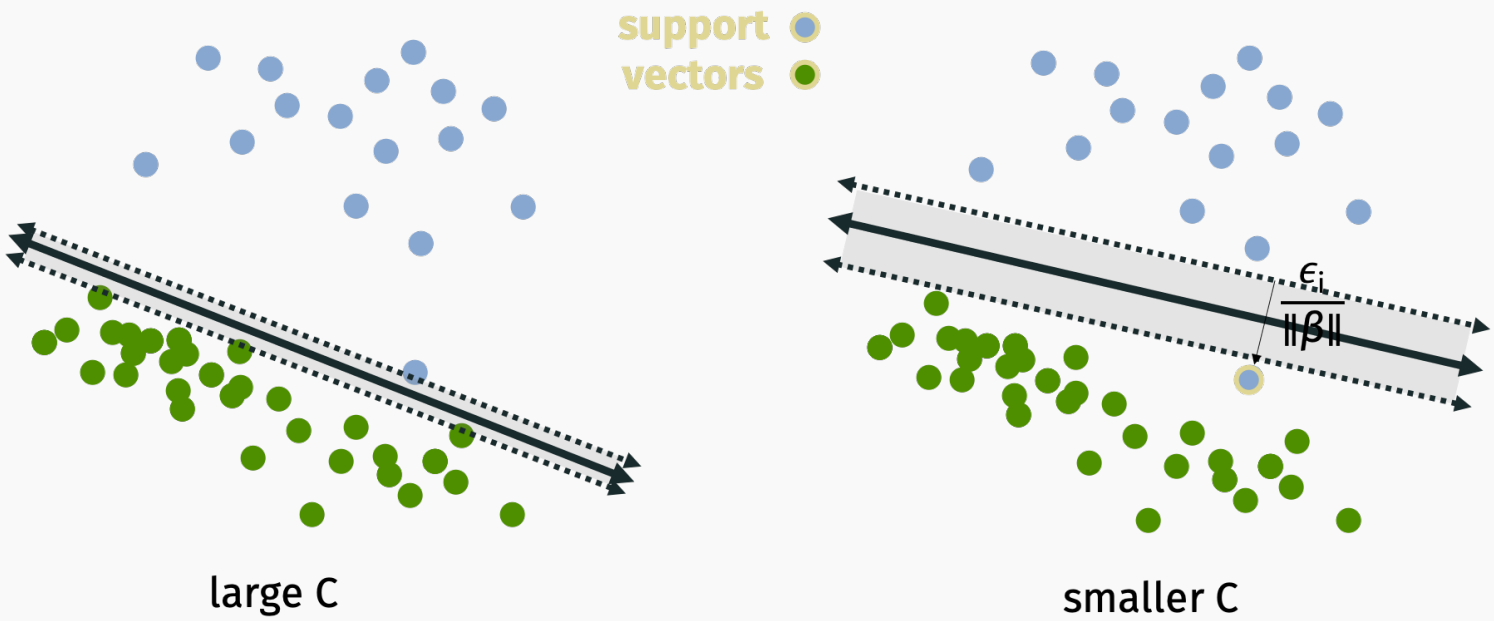
When data is linearly separable, as $C \to \infty$ we will always get a separating hyperplane. A smaller value of C might lead to a more robust solution.

EFFECT OF C

Example dataset:



EFFECT OF C



The classifier on the right is intuitively more robust. So for this data, a smaller choice for C might make sense.

Some basic transformations of the soft-margin objective:

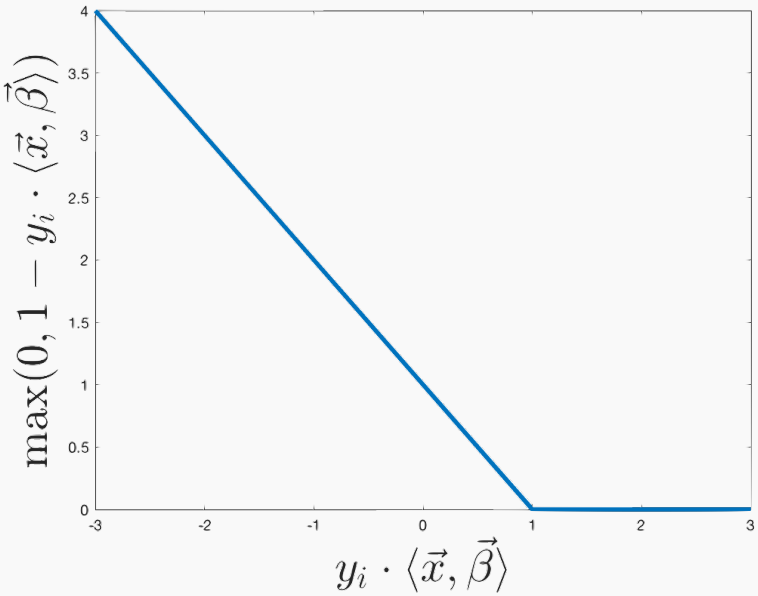
$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_2^2 + C \sum_{i=1}^{n} \epsilon_i \quad \text{subject to} \quad y_i \cdot \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle \geq 1 - \epsilon_i \text{ for all } i.$$

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_2^2 + C \sum_{i=1}^{n} \max(0, 1 - y_i \cdot \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle).$$

$$\min_{\boldsymbol{\beta}} \lambda \|\boldsymbol{\beta}\|_2^2 + \sum_{i=1}^{n} \max(0, 1 - y_i \cdot \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle).$$

These are all equivalent. $\lambda = 1/C$ is just another scaling parameter.

Hinge-loss: $\max(0, 1 - y_i \cdot \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle)$. Recall that $y_i \in \{-1, 1\}$.



Soft-margin SVM:

$$\min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n} \max(0, 1 - y_i \cdot \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle) + \lambda \|\boldsymbol{\beta}\|_2^2 \right]. \tag{1}$$

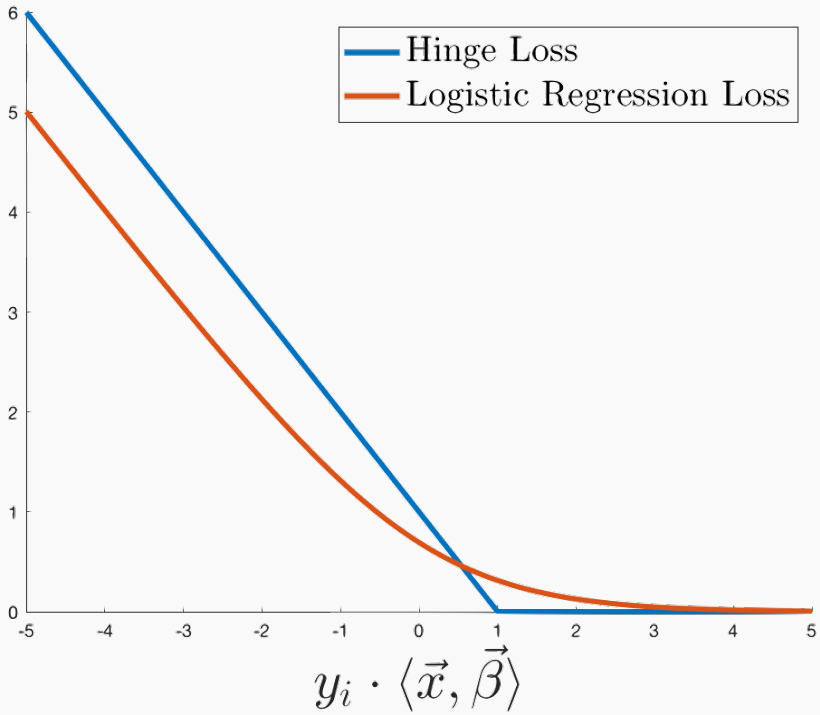
Recall the logistic loss for $y_i \in \{0, 1\}$:

$$
\begin{aligned}
L(\boldsymbol{\beta}) &= -\sum_{i=1}^{n} y_i \log(h(\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle)) + (1 - y_i) \log(1 - h(\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle)) \\
&= -\sum_{i=1}^{n} y_i \log\left(\frac{1}{1 + e^{-\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle}}\right) + (1 - y_i) \log\left(\frac{e^{-\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle}}{1 + e^{-\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle}}\right) \\
&= -\sum_{i=1}^{n} y_i \log\left(\frac{1}{1 + e^{-\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle}}\right) + (1 - y_i) \log\left(\frac{1}{1 + e^{\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle}}\right)
\end{aligned}
$$

comparison of SVM to logistic regression

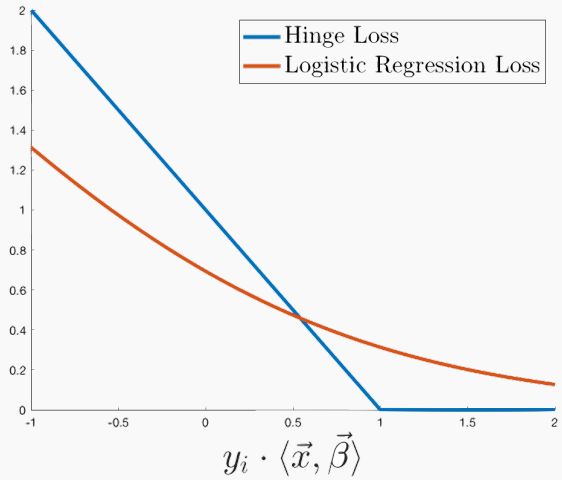
Compare this to the logistic regression loss reformulated for $y_i \in \{-1, 1\}$):

$$\sum_{i=1}^{n} -\log\left(\frac{1}{1 - e^{-y_i \cdot \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle}}\right)$$



COMPARISON TO LOGISTIC REGRESSION

So, in the end, the function minimized when finding β for the standard **soft-margin SVM** is very similar to the objective function minimized when finding β using **logistic regression with ℓ_2 regularization.** Sort of...



Both functions can be optimized using first-order methods like gradient descent. This is now a common choice for large problems. **Will explore more on next lab.**