Machine Learning

A. G. Schwing & M. Telgarsky

University of Illinois at Urbana-Champaign, 2018

L6: Support Vector Machines (SVMs)

Note to those reading at home:
stuff is derived on the board, not in these slides.

Lecture outline.

1. Review.
2. Motivation in separable case.
3. Nonseparable case.
4. Duality, support vectors, kernels.
5. Odds and ends.

Reading.

- K. Murphy; Machine Learning: A Probabilistic Perspective; Chapter 14.5.

Review.

Review.

Lectures so far:

1. Basic ML; k-nn (k nearest neighbor).
2. Least squares (linear regression).
3. Logistic regression.
4. Convexity and optimization I.
5. Convexity and optimization II.
6. **Support vector machines.**

Review.

We have two ways to learn linear predictors (via ERM):

- **Least squares:**

$$\arg\min_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}(y^{(i)} - \boldsymbol{w}^\top \mathbf{x}^{(i)})^2.$$

- **Logistic regression:**

$$\arg\min_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}\ln\left(1 + \exp(-y^{(i)}\boldsymbol{w}^\top \mathbf{x}^{(i)})\right).$$

We said logistic regression is better for classification ($y \in \{-1, +1\}$).

Review.

We have two ways to learn linear predictors (via ERM):

- **Least squares:**

$$\arg\min_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}(y^{(i)} - \boldsymbol{w}^\top \mathbf{x}^{(i)})^2.$$

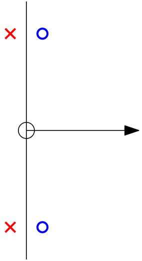
- **Logistic regression:**

$$\arg\min_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \ln\left(1 + \exp(-y^{(i)}\boldsymbol{w}^\top \mathbf{x}^{(i)})\right).$$

We said logistic regression is better for classification ($y \in \{-1, +1\}$).

Can we build a classifier *explicitly* for good classification?

Linear classifiers.

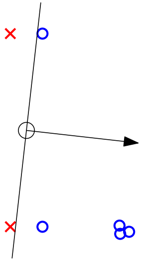


Consider finding $\boldsymbol{w} \in \mathbb{R}^2$ with ERM, meaning

$$\arg\min_{\boldsymbol{w} \in \mathbb{R}^2} \frac{1}{n} \sum_{i=1}^{n} \ell\left(y^{(i)} \boldsymbol{w}^\top \mathbf{x}^{(i)}\right),$$

where ℓ is **convex** and cares about magnitude of error.

Linear classifiers.



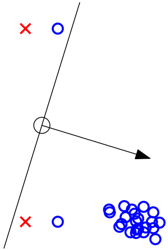
Consider finding $\boldsymbol{w} \in \mathbb{R}^2$ with ERM, meaning

$$\underset{\boldsymbol{w} \in \mathbb{R}^2}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \ell \left(y^{(i)} \boldsymbol{w}^\top \mathbf{x}^{(i)} \right),$$

where ℓ is **convex** and cares about magnitude of error.

Adding points...

Linear classifiers.



Consider finding $\boldsymbol{w} \in \mathbb{R}^2$ with ERM, meaning

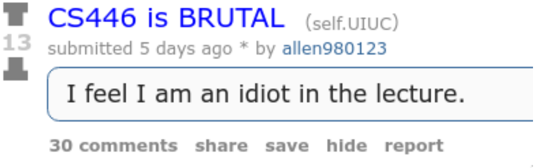
$$\operatorname*{arg\,min}_{\boldsymbol{w}\in\mathbb{R}^2} \frac{1}{n}\sum_{i=1}^{n} \ell\left(y^{(i)}\boldsymbol{w}^\top \mathbf{x}^{(i)}\right),$$

where ℓ is **convex** and cares about magnitude of error.

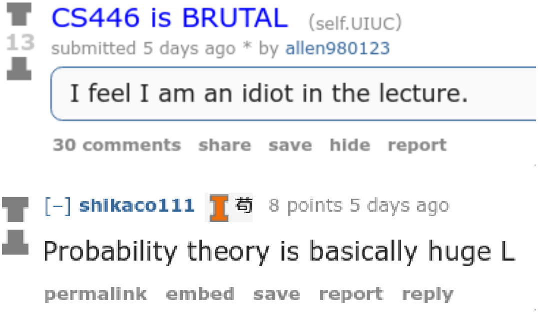
Adding points...
...causes logistic regression and least squares to misclassify!

End of review; an aside.

End of review; an aside.



End of review; an aside.



CS446 is BRUTAL (self.UIUC)

13

submitted 5 days ago * by allen980123

I feel I am an idiot in the lecture.

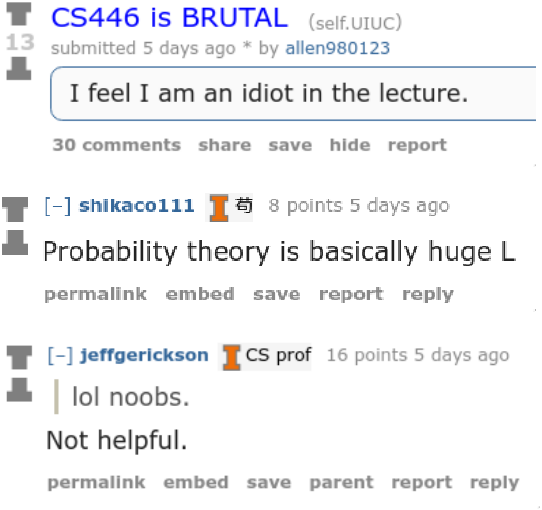
30 comments share save hide report

[–] shikaco111 苟 8 points 5 days ago

Probability theory is basically huge L

permalink embed save report reply

End of review; an aside.



End of review; an aside.

[–] -mjt- 12 points 5 days ago

Hi,

Matus here. I'll try to get this account verified. What feedback would you like to give me?

I received a comment that my boardwork was hard to read.

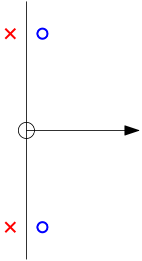
Also, sleep deprivation meant I had to make one joke (that only I found funny) per minute.

permalink embed save report reply

> [–] allen980123 [S] 8 points 5 days ago
>
> Personally I think more intuitive explanation / graphs on formulas/proofs would help greatly. The pace of the class is a bit fast. I know it's impossible to explain everything well in 75 minutes and as you've heard people can't read the blackboard clearly. Maybe put them on slides or make them as a separate note as reference would be a great idea.
>
> A lot of people actually don't have much Math background beyond calculus/linalg. Some notations, for example sup and inf are new to many people and make the formula difficult to understand.
>
> This is just my opinion so please ask more people.
>
> I very appreciate everything you and Alex do to make this class great.
>
> Also are we going to proof those formulas/inequalities in exams?
>
> permalink embed save parent report reply
>
> [–] mtgross12 1 point 3 days ago
>
> ProTip: Make notes of what you want to cover in class before each lecture and scan them in and post them online so the class can follow along / review after.
>
> I personally love when professors do this because then I can use the weekends to review lectures and compress notes onto a study guide that I can use when exam time comes around.
>
> permalink embed save parent report reply

SVM — motivation in separable case.

Linear classifiers.

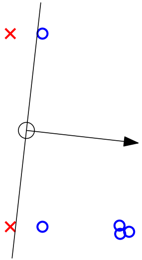


Consider finding $\boldsymbol{w} \in \mathbb{R}^2$ with ERM, meaning

$$\arg\min_{\boldsymbol{w} \in \mathbb{R}^2} \frac{1}{n} \sum_{i=1}^{n} \ell\left(y^{(i)} \boldsymbol{w}^\top \mathbf{x}^{(i)}\right),$$

where ℓ is **convex** and cares about magnitude of error.

Linear classifiers.



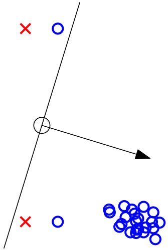
Consider finding $\boldsymbol{w} \in \mathbb{R}^2$ with ERM, meaning

$$\arg\min_{\boldsymbol{w} \in \mathbb{R}^2} \frac{1}{n} \sum_{i=1}^{n} \ell\left(y^{(i)} \boldsymbol{w}^\top \mathbf{x}^{(i)}\right),$$

where ℓ is **convex** and cares about magnitude of error.

Adding points. . .

Linear classifiers.



Consider finding $\boldsymbol{w} \in \mathbb{R}^2$ with ERM, meaning

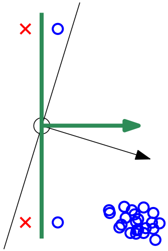
$$\underset{\boldsymbol{w}\in\mathbb{R}^2}{\arg\min}\ \frac{1}{n}\sum_{i=1}^{n}\ell\left(y^{(i)}\boldsymbol{w}^\top\mathbf{x}^{(i)}\right),$$

where ℓ is **convex** and cares about magnitude of error.

Adding points. . .

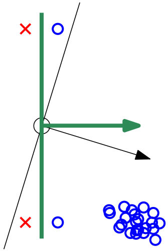
. . . causes logistic regression and least squares to misclassify!

Linear classifier via linear programming.



How to pick $\boldsymbol{w} \in \mathbb{R}^2$ so that all predictions correct?

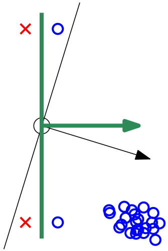
Linear classifier via linear programming.



How to pick $\boldsymbol{w} \in \mathbb{R}^2$ so that all predictions correct?

$$\text{Find } \boldsymbol{w} \in \mathbb{R}^2 \qquad \text{s.t. } y^{(i)} \boldsymbol{w}^\top \mathbf{x}^{(i)} > 0 \quad \forall i \in \{1, \ldots, n\}\,.$$

Linear classifier via linear programming.

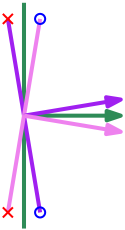


How to pick $\boldsymbol{w} \in \mathbb{R}^2$ so that all predictions correct?

$$\text{Find } \boldsymbol{w} \in \mathbb{R}^2 \qquad \text{s.t. } y^{(i)} \boldsymbol{w}^\top \mathbf{x}^{(i)} > 0 \quad \forall i \in \{1, \ldots, n\}\,.$$

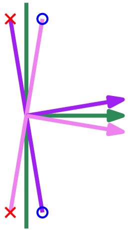
This is a **linear feasibility problem**, thus solvable (when feasible).

Linear classifiers and maximum margins.



Question: which (correct) classifier?

Linear classifiers and maximum margins.

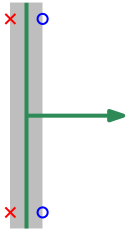


Question: which (correct) classifier?

Maximum margin principle (Vapnik, '82)**:**

> *Choose* $\boldsymbol{w} \in \mathbb{R}^2$ *which maximizes* ***margin*** *(distance to closest data point).*

Linear classifiers and maximum margins.



Question: which (correct) classifier?

Maximum margin principle (Vapnik, '82)**:**

> *Choose* $\boldsymbol{w} \in \mathbb{R}^2$ *which maximizes* ***margin***
> *(distance to closest data point).*

Maximum margins.

Maximize margin, meaning distance to closest example.

Maximum margins.

Maximize margin, meaning distance to closest example.

Given $\boldsymbol{w}$, distance to closest example is

Maximum margins.

Maximize margin, meaning distance to closest example.

Given $\boldsymbol{w}$, distance to closest example is

$$\min_{1 \le i \le n} \frac{y^{(i)} \boldsymbol{w}^\top \mathbf{x}^{(i)}}{\|\boldsymbol{w}\|_2}.$$

Maximum margin classifier given by

$$\max_{\boldsymbol{w} \in \mathbb{R}^d} \min_{1 \le i \le n} \frac{y^{(i)} \boldsymbol{w}^\top \mathbf{x}^{(i)}}{\|\boldsymbol{w}\|_2}.$$

Maximum margins.

Maximize margin, meaning distance to closest example.
Given $\boldsymbol{w}$, distance to closest example is

$$\min_{1 \leq i \leq n} \frac{y^{(i)} \boldsymbol{w}^\top \mathbf{x}^{(i)}}{\|\boldsymbol{w}\|_2}.$$

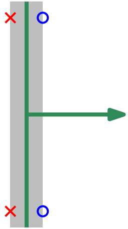
Maximum margin classifier given by

$$\max_{\boldsymbol{w} \in \mathbb{R}^d} \min_{1 \leq i \leq n} \frac{y^{(i)} \boldsymbol{w}^\top \mathbf{x}^{(i)}}{\|\boldsymbol{w}\|_2}.$$

Simplification: introduce constraints:

$$\begin{aligned}
&\max_{\boldsymbol{w} \in \mathbb{R}^d, r \geq 0} \frac{r}{\|\boldsymbol{w}\|_2} && \text{s.t.} \quad r \leq y^{(i)} \boldsymbol{w}^\top \mathbf{x}^{(i)} \quad \forall i \in \{1, \ldots, n\} \\
= &\max_{\boldsymbol{w} \in \mathbb{R}^d, r \geq 0} \frac{1}{\|\boldsymbol{w}/r\|_2} && \text{s.t.} \quad 1 \leq y^{(i)} (\boldsymbol{w}/r)^\top \mathbf{x}^{(i)} \quad \forall i \in \{1, \ldots, n\} \\
= &\max_{\boldsymbol{w} \in \mathbb{R}^d} \frac{1}{\|\boldsymbol{w}\|_2} && \text{s.t.} \quad 1 \leq y^{(i)} \boldsymbol{w}^\top \mathbf{x}^{(i)} \quad \forall i \in \{1, \ldots, n\}.
\end{aligned}$$

Maximum margins linear classifier.



Find the separator which **maximizes margin**:

$$\arg\min_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{2}\|\boldsymbol{w}\|_2^2 \qquad \text{s.t.} \quad 1 \leq y^{(i)}\boldsymbol{w}^\top\mathbf{x}^{(i)} \quad \forall i \in \{1,\ldots,n\}\,.$$

This optimization problem:

- is *convex*;
- if a solution exists, it is unique.

SVM dual problem.

Primal:

$$P(w) := \begin{cases} \frac{1}{2}\|\boldsymbol{w}\|_2^2 & \text{when } 1 \leq y^{(i)}\boldsymbol{w}^\top \mathbf{x}^{(i)} \quad \forall i \in \{1, \ldots, n\}\,; \\ \infty & \text{otherwise.} \end{cases}$$

Lagrangian (with Lagrange multipliers $\alpha \geq 0$)**:**

$$L(\boldsymbol{w}, \alpha) := \frac{1}{2}\|w\|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - y^{(i)}\boldsymbol{w}^\top \mathbf{x}^{(i)});$$

note $P(\boldsymbol{w}) = \sup_{\alpha \geq 0} L(\boldsymbol{w}, \alpha)$. **Dual:**

$$D(\alpha) := \inf_{\boldsymbol{w} \in \mathbb{R}^d} L(\boldsymbol{w}, \alpha) = \begin{cases} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\left\|\sum_{i=1}^{n} \alpha_i y^{(i)} \mathbf{x}^{(i)}\right\|^2 & \alpha \geq 0, \\ -\infty & \text{otherwise.} \end{cases}$$

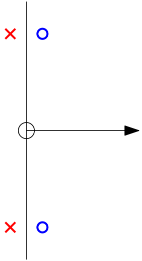
Nonseparable case.

Nonseparable case.

Recall the original **linear feasibility problem**:

$$\text{find } \boldsymbol{w} \in \mathbb{R}^d \qquad \text{s.t.} \quad y^{(i)} \boldsymbol{w}^\top \mathbf{x}^{(i)} > 0 \quad \forall i \in \{1, \ldots, n\}.$$

What does “infeasible” mean geometrically?



Nonseparable case; a relaxed program.

We can add **slack variables** into the feasibility program:

$$\min_{\boldsymbol{w}\in\mathbb{R}^2} \quad 0 \qquad\qquad \text{s.t.} \quad y^{(i)}\boldsymbol{w}^\top \mathbf{x}^{(i)} > 0 \qquad \forall i \in \{1, \ldots, n\}\,.$$

Geometric interpretation:
$\sum_i \xi_i$ is minimal translation to get feasible problem.

Technical note: open constraint for discussion only...

Nonseparable case; a relaxed program.

We can add **slack variables** into the feasibility program:

$$\min_{\boldsymbol{w}\in\mathbb{R}^2, \boldsymbol{\xi}\in\mathbb{R}^n_{\geq 0}} 0 + \sum_{i=1}^{n} \boldsymbol{\xi}_i \qquad \text{s.t.} \qquad y^{(i)}\boldsymbol{w}^\top \mathbf{x}^{(i)} > 0 - \boldsymbol{\xi}_i \quad \forall i \in \{1, \ldots, n\}.$$

Geometric interpretation:
$\sum_i \xi_i$ is minimal translation to get feasible problem.

Technical note: open constraint for discussion only...

Maximum margin solution in nonseparable case.

We can also add **slack variables** to the maximum margin program:

$$\min_{\boldsymbol{w}\in\mathbb{R}^d,\boldsymbol{\xi}\in\mathbb{R}^n_{\geq 0}} \frac{1}{2}\|w\|_2^2+C\sum_{i=1}^{n}\xi_i \quad \text{s.t.} \quad 1-\xi_i \leq y^{(i)}\boldsymbol{w}^\top\mathbf{x}^{(i)} \quad \forall i \in \{1,\ldots,n\}\,.$$

This is sometimes called **soft-margin SVM**.

Maximum margin solution in nonseparable case.

We can also add **slack variables** to the maximum margin program:

$$\min_{\boldsymbol{w}\in\mathbb{R}^d,\boldsymbol{\xi}\in\mathbb{R}^n_{\geq 0}} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n}\xi_i \quad \text{s.t.} \quad 1-\xi_i \leq y^{(i)}\boldsymbol{w}^\top\mathbf{x}^{(i)} \quad \forall i\in\{1,\ldots,n\}\,.$$

This is sometimes called **soft-margin SVM**.

Question: why "C"?

Maximum margin solution in nonseparable case — other forms.

Original:

$$\min_{\boldsymbol{w}\in\mathbb{R}^d,\boldsymbol{\xi}\in\mathbb{R}^n_{\geq 0}} \frac{1}{2}\|w\|_2^2+C\sum_{i=1}^n \xi_i \quad \text{s.t.} \quad 1-\xi_i \leq y^{(i)}\boldsymbol{w}^\top\mathbf{x}^{(i)} \quad \forall i\in\{1,\ldots,n\}.$$

Regularized form:

$$\min_{\boldsymbol{w}\in\mathbb{R}^d,\boldsymbol{\xi}\in\mathbb{R}^n_{\geq 0}} \sum_{i=1}^n \xi_i + \frac{\lambda}{2}\|w\|_2^2 \quad \text{s.t.} \quad 1-\xi_i \leq y^{(i)}\boldsymbol{w}^\top\mathbf{x}^{(i)} \quad \forall i\in\{1,\ldots,n\}.$$

Unconstrained form:

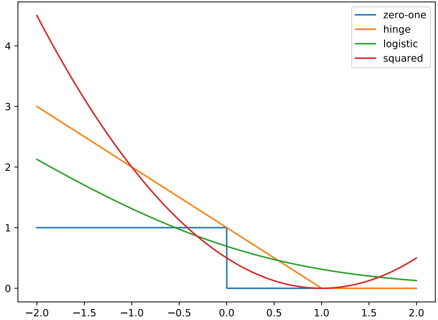
$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \sum_{i=1}^n \ell_{\text{hinge}}(y^{(i)}\boldsymbol{w}^\top\mathbf{x}^{(i)}) + \frac{\lambda}{2}\|w\|^2 \quad \text{where } \ell_{\text{hinge}}(z)=\max\{0,1-z\}.$$

Last one is what most people call **Support Vector Machine (SVM).**

Comparison of losses.

Unconstrained form:

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \sum_{i=1}^{n} \ell_{\text{hinge}}(y^{(i)}\boldsymbol{w}^\top \mathbf{x}^{(i)}) + \frac{\lambda}{2}\|w\|^2 \quad \text{where } \ell_{\text{hinge}}(z) = \max\{0, 1-z\}$$



Remark. Which loss? See “Statistical behavior and consistency of classification methods based on convex risk minimization”, Zhang 2004.

Duality, support vectors, kernels.

Dual of slack formultion.

Primal:

$$P(\boldsymbol{w}, \xi) := \begin{cases} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^n \xi_i & 1 - \xi_i \leq y^{(i)}\boldsymbol{w}^\top \mathbf{x}^{(i)} \quad \forall i \in \{1, \ldots, n\}, \\ \infty & \text{otherwise.} \end{cases}$$

Lagrangian (with $\alpha \geq 0$)**:**

$$L(\boldsymbol{w}, \xi, \alpha) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^n \xi_i + \sum_{i=1}^n \alpha_i(1 - \xi_i - y^{(i)}\boldsymbol{w}^\top \mathbf{x}^{(i)}).$$

Dual (derived as $\sup_{\boldsymbol{w},\xi} L(\boldsymbol{w}, \xi, \alpha)$)**:**

$$D(\alpha) = \begin{cases} \sum_{i=1}^n \alpha_i - \frac{1}{2}\left\|\sum_{i=1}^n \alpha_i y^{(i)}\mathbf{x}^{(i)}\right\|^2 & 0 \leq \alpha_i \leq C; \\ -\infty & \text{otherwise.} \end{cases}$$

Remark. Some literature has a different dual, due to threshold.

Support vectors.

Dual program

$$\max_{\alpha \in [0,C]^n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \left\| \sum_{i=1}^{n} \alpha_i y^{(i)} \mathbf{x}^{(i)} \right\|^2 .$$

Support vectors.

Dual program

$$\max_{\alpha \in [0,C]^n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \left\| \sum_{i=1}^{n} \alpha_i y^{(i)} \mathbf{x}^{(i)} \right\|^2 .$$

Derivation gave

$$\boldsymbol{w} := \sum_i \alpha_i y^{(i)} \mathbf{x}^{(i)},$$

which only depends on $(\boldsymbol{x}^{(i)}, y^{(i)})$ *with* $\alpha_i > 0$*.*

These examples are **support vectors.**

Can throw away other examples and solution unchanged.

Support vectors.

Dual program

$$\max_{\alpha \in [0,C]^n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \left\| \sum_{i=1}^{n} \alpha_i y^{(i)} \mathbf{x}^{(i)} \right\|^2 .$$

Derivation gave

$$\boldsymbol{w} := \sum_i \alpha_i y^{(i)} \mathbf{x}^{(i)},$$

which only depends on $(\boldsymbol{x}^{(i)}, y^{(i)})$ *with* $\alpha_i > 0$*.*

These examples are **support vectors.**

Can throw away other examples and solution unchanged.

Question: geometric meaning?

Kernels.

Suppose $\mathbf{x}^{(i)}$ replaced with $\phi(\mathbf{x}^{(i)})$:

$$\max_{\boldsymbol{\alpha}\in[0,C]^n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\left\|\sum_{i=1}^{n} \alpha_i y^{(i)} \phi(\mathbf{x}^{(i)})\right\|^2$$

$$= \max_{\boldsymbol{\alpha}\in[0,C]^n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} (\phi(\mathbf{x}^{(i)}))^\top \phi(\mathbf{x}^{(j)}).$$

Kernels.

Suppose $\mathbf{x}^{(i)}$ replaced with $\phi(\mathbf{x}^{(i)})$:

$$\max_{\boldsymbol{\alpha}\in[0,C]^n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\left\|\sum_{i=1}^{n} \alpha_i y^{(i)} \phi(\mathbf{x}^{(i)})\right\|^2$$
$$= \max_{\boldsymbol{\alpha}\in[0,C]^n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} (\phi(\mathbf{x}^{(i)}))^\top \phi(\mathbf{x}^{(j)}).$$

Replace $\phi(\boldsymbol{x})^\top \phi(\boldsymbol{x}')$ with $k(\boldsymbol{x}, \boldsymbol{x}')$ for some **kernel function** k;
$\phi(\boldsymbol{x})$ becomes implicit!

Kernels.

Suppose $\mathbf{x}^{(i)}$ replaced with $\phi(\mathbf{x}^{(i)})$:

$$\begin{aligned} &\max_{\boldsymbol{\alpha}\in[0,C]^n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\left\|\sum_{i=1}^{n} \alpha_i y^{(i)} \phi(\mathbf{x}^{(i)})\right\|^2 \\ &= \max_{\boldsymbol{\alpha}\in[0,C]^n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} (\phi(\mathbf{x}^{(i)}))^\top \phi(\mathbf{x}^{(j)}). \end{aligned}$$

Replace $\phi(\boldsymbol{x})^\top \phi(\boldsymbol{x}')$ with $k(\boldsymbol{x}, \boldsymbol{x}')$ for some **kernel function** k; $\phi(\boldsymbol{x})$ becomes implicit!

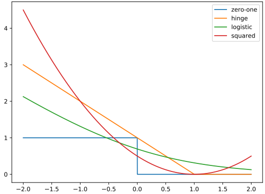
At prediction time:

$$\boldsymbol{x} \mapsto \sum_{i=1}^{n} \alpha_i y^{(i)} k(\mathbf{x}^{(i)}, \boldsymbol{x}).$$

Odds and ends.

Hinge loss?

Recall the hinge loss $\ell_{\text{hinge}}(z) := \max\{0, 1 - z\}$.



For any vector $\boldsymbol{v}$:

$$0 \leq \frac{1}{r} \ln \sum_{i=1}^{n} \exp(r \boldsymbol{v}_i) - \|v\|_\infty \leq \frac{\ln(n)}{r}.$$

Thus logistic and hinge related:

$$\lim_{r \to \infty} \ln(1 + \exp(-r \cdot z)) = \max\{0, -z\}.$$

SVM and SGD.

Suppose we get $(\boldsymbol{x}, y)$; what's our stochastic gradient?

SVM and SGD.

Suppose we get $(\boldsymbol{x}, y)$; what's our stochastic gradient?

The stochastic gradient update for $\ell_{\text{hinge}}(y\boldsymbol{w}^\top\boldsymbol{x}) + \lambda\|w\|^2/2$ is

$$\boldsymbol{w}' := (1-\lambda)\boldsymbol{w} + y\boldsymbol{x} \cdot \mathbb{1}[y\boldsymbol{w}^\top\boldsymbol{x} < 1].$$

Geometric view?

SVM and SGD.

Suppose we get $(\boldsymbol{x}, y)$; what's our stochastic gradient?

The stochastic gradient update for $\ell_{\text{hinge}}(y\boldsymbol{w}^\top\boldsymbol{x}) + \lambda\|w\|^2/2$ is

$$\boldsymbol{w}' := (1-\lambda)\boldsymbol{w} + y\boldsymbol{x} \cdot \mathbb{1}[y\boldsymbol{w}^\top\boldsymbol{x} < 1].$$

Geometric view? Rotate towards margin violations;
keep predictor small.

SVM and SGD.

Suppose we get $(\boldsymbol{x}, y)$; what's our stochastic gradient?

The stochastic gradient update for $\ell_{\text{hinge}}(y\boldsymbol{w}^\top\boldsymbol{x}) + \lambda\|w\|^2/2$ is

$$\boldsymbol{w}' := (1-\lambda)\boldsymbol{w} + y\boldsymbol{x} \cdot \mathbb{1}[y\boldsymbol{w}^\top\boldsymbol{x} < 1].$$

Geometric view? Rotate towards margin violations; keep predictor small.

Note. Can also do some projection; google "pegasos".

Summary and key concepts.

- Exact linear classifier via linear programming (when separable!).
- Maximum margin classifiers.
- SVM.
- Hinge loss.
- SVM dual.

Dual derivation hints — separable case.

For the separable case problem, note

$$0 = \nabla_{\boldsymbol{w}} L(\boldsymbol{w}, \alpha) = \boldsymbol{w} - \sum_{i=1}^{n} \alpha_i y^{(i)} \mathbf{x}^{(i)},$$

and plugging $\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y^{(i)} \mathbf{x}^{(i)}$ into $\mathcal{L}(\boldsymbol{w}, \alpha)$ and collecting terms gives the stated expression for $D(\alpha)$.

Dual derivation hints — nonseparable case.

For the nonseparable case, there are both $\boldsymbol{w}$ and ξ to worry about. Optimizing $\boldsymbol{w}$ proceeds exactly as before:

$$0 = \nabla_{\boldsymbol{w}} L(\boldsymbol{w}, \boldsymbol{\alpha}) = \boldsymbol{w} - \sum_{i=1}^{n} \alpha_i y^{(i)} \mathbf{x}^{(i)},$$

which suggests $\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y^{(i)} \mathbf{x}^{(i)}$. To optimize $\boldsymbol{\xi}$, a derivative gives nothing, but isolating the terms with $\boldsymbol{\xi}$ gives

$$\sup_{\boldsymbol{\xi}_i \geq 0} \boldsymbol{\xi}_i (C - \boldsymbol{\alpha}_i);$$

if $C > \boldsymbol{\alpha}_i$, then this expression becomes $+\infty$, which implies a constraint $\boldsymbol{\alpha}_i \leq C$, in which case this expression is 0.