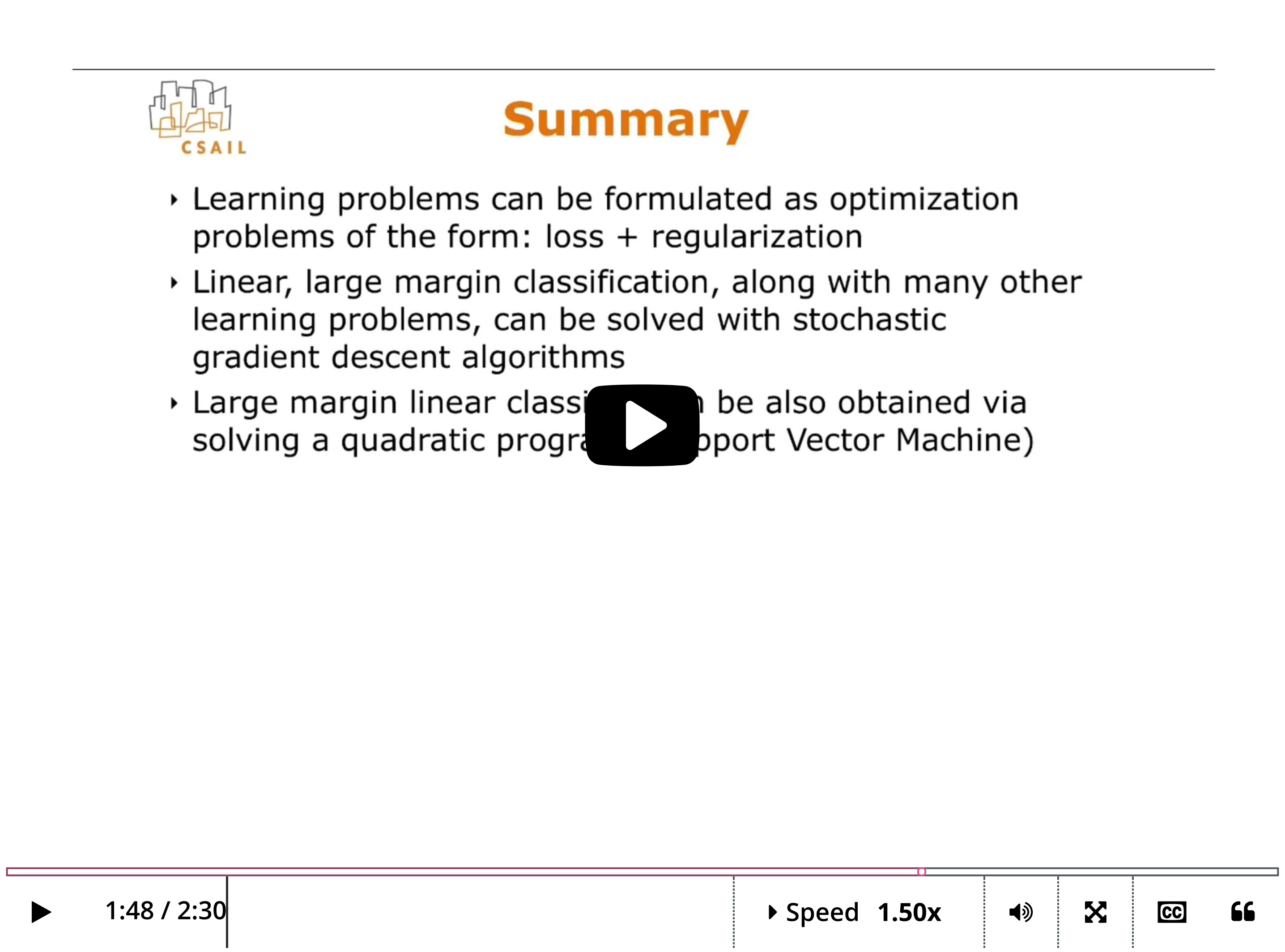
6. The Realizable Case - Quadratic program

The Realizable Case - Quadratic program



Summary

- Learning problems can be formulated as optimization problems of the form: loss + regularization
- Linear, large margin classification, along with many other learning problems, can be solved with stochastic gradient descent algorithms
- Large margin linear classi be also obtained via solving a quadratic progra pport Vector Machine)

because we increase the margin until we cannot do it any more,

and we start hitting the training samples.

But we cannot go further because in this simple case,

we strictly try to enforce the margin constraints.

What we have seen so far is how to understand the optimization

problem corresponding to the maximum margin

linear classification, the effect of regularization as we

change the regularization parameter, how

the solution changes qualitatively

as well as in terms of generalization.

We also saw how to actually solve

the associated optimization problem using gradient descent

updates, in particular stochastic gradient descent

updates that I present to perform for such functions.

We also briefly discussed how to turn

Video

Download video file

Transcripts

Download SubRip (.srt) file
Download Text (.txt) file

The realizable case 1

1/1 point (graded)
In the realizable case, which of the following is true?

○ There is exactly one (θ, θ_0) that satisfies $y^{(i)}\left(\theta \cdot x^{(i)} + \theta_0\right) >= 1$ for $i = 1, \ldots n$.

○ There are more than one, but finite number of (θ, θ_0) that satisfy $y^{(i)}\left(\theta \cdot x^{(i)} + \theta_0\right) >= 1$ for $i = 1, \ldots n$.

● There are infinitely many (θ, θ_0) that satisfy $y^{(i)}\left(\theta \cdot x^{(i)} + \theta_0\right) >= 1$ for $i = 1, \ldots n$. ✓

Solution:

Without any additional constraint, because θ and θ_0 are continuous, there are numerously many (θ, θ_0) that satisfy the zero-error case.

Submit You have used 2 of 2 attempts

Answers are displayed within the problem

The realizable case 2

1/1 point (graded)
Remember the objective function

$$J(\theta, \theta_0) = \frac{1}{n} \sum_{i=1}^{n} \text{Loss}_h\left(y^{(i)}\left(\theta \cdot x^{(i)} + \theta_0\right)\right) + \frac{\lambda}{2} \mid\mid \theta \mid\mid^2$$

In the realizable case, we can always find (θ, θ_0) such that the sum of the hinge losses is 0. In this case, what does the objective function J reduce to?

○ $\frac{1}{n} \sum_{i=1}^{n} \text{Loss}_h\left(y^{(i)}\left(\theta \cdot x^{(i)} + \theta_0\right)\right)$

○ $\frac{1}{n}\sum_{i=1}^{n}\text{Loss}_h\left(y^{(i)}\left(\theta\cdot x^{(i)}+\theta_0\right)\right)+\frac{\lambda}{2}\mid\mid\theta\mid\mid^2$

● $\frac{1}{2}\mid\mid\theta\mid\mid^2$ ✓

Solution:

In the realizable case, we can always find a decision boundary such that the first term of $J\left(\theta,\theta_0\right)$ is 0. Thus $J\left(\theta,\theta_0\right)$ reduces to $\frac{\lambda}{2}\mid\mid\theta\mid\mid^2$. Our goal is to find θ that minimizes J anyways, so J reduces to $\frac{1}{2}\mid\mid\theta\mid\mid^2$

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

Support Vectors

1/1 point (graded)

Support vectors refer to points that are exactly on the margin boundary. Which of the following is true? Choose all those apply.

□ If we remove one point that is not a support vector, we will get a different θ,θ_0

☑ If we remove all points that are support vectors, we will get a different θ,θ_0 ✓

□ If we remove one point that is a support vector, we will get the same θ,θ_0

☑ If we remove one point that is not a support vector, we will get the same θ,θ_0 ✓

✓

Solution:

Support vectors determine the exact solution θ,θ_0 that minimizes $J\left(\theta,\theta_0\right)$. Thus removing/changing all of them changes the θ,θ_0. On the other hand, any training example that is not a support vector has no influence on θ,θ_0. Thus removing/changing them does not affect θ,θ_0.

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

Discussion

Hide Discussion

Topic: Unit 1 Linear Classifiers and Generalizations (2 weeks):Lecture 4. Linear Classification and Generalization / 6. The Realizable Case - Quadratic program

Add a Post

All Posts

The realizable case 2

discussion posted 7 days ago by NSchandler

I believe that this question is enforcing an unstated restriction that we choose ex ante to have no margin violations.

In general, just because the data is linearly separable doesn't mean J is minimized at a point with no margin violations.

This post is visible to everyone.

Add a Response

4 responses

mrBB (Community TA)
7 days ago

I think you bring up a valid point. I now think I read the question too casual and received an undeserved green tick.

Add a comment

Mark_B2 (Community TA)
7 days ago

The question is not about minimizing J. It asks, assuming it's possible to find (θ, θ_0) that separates the data with maximal possible margin, what would be a value of J. Would it be the minimum J? Not necessarily, depends on λ.

I see what you're saying, but not totally convinced that's what the question is literally asking. Or at least, it seems ambiguous what the question is asking.

Relevant portion of the question:

> In the realizable case, we can always find (θ, θ_0) such that the sum of the hinge losses is 0. In this case, what does the objective function J reduce to?

I read that essentially as "In the realizable case, what does the objective function J reduce to?" That is, "this case" refers to "the realizable case."

I believe you are reading it such that "this case" refers to the 2nd half of the prior sentence, so the question asks "if the sum of the hinge losses is 0, what does the objective function J reduce to."

I suppose I can see both readings. I still believe that the question is literally asking "what does the objective function reduce to in the realizable case," but perhaps I'm wrong. Maybe our resident grammar expert ptressel can chime in :)

posted 7 days ago by **NSchandler**

My reading now is that we have to assume θ, θ_0 exist that make the Hinge Loss zero. But that is a different statement than that we have to assume that the Hinge Loss is zero. So I read "in this case" as "θ, θ_0 exist that make the Hinge Loss zero" and not as "Hinge Loss is zero" or "assume Hinge Loss is zero".

posted 7 days ago by **mrBB** (Community TA)

This is all in the context of converting a linearly separable problem to something one can toss at a quadratic programming solver. It's the solver's job to maintain the constraints, and then optimize an objective subject to those constraints. Since the solver is dealing with satisfying the constraints under the covers (so the hinge loss *will* be zero), what objective function should we give it to minimize?

posted 6 days ago by **ptressel** (Community TA)

Add a comment

ptressel (Community TA)
6 days ago

Yes, with a tunable λ in $J(\theta, \theta_0)$, then we can bork (https://www.merriam-webster.com/dictionary/bork 2nd definition) the nice empty margin by cranking λ up. And yes, we might want to do that even if the points are separable -- that allows choosing between basing the boundary on just the outermost points, which could be noisy, or basing it more on where the bulk of the points are.

(Wildly off topic, but...hmm. The MW dictionary is very very wrong about the seniority of their 1st definition of bork. The 2nd meaning is *much* older than 2003. There's an Urban Dictionary entry dating from 2001, which all by itself denies their dating. The folks updating dictionaries may not have heard the term, so why not just ask us geeks when it first appeared? We can probably find instances in Usenet newsgroups... Aaaand, it's in the

Jargon File: http://catb.org/jargon/html/B/borken.html .)

Add a comment

mbh038

3 days ago

I am confused about this question. If $J\left(\theta, \theta_0\right)$ reduces to $||\ \theta\ ||^2$, don't we end up with $\theta = 0$??

No, θ must also satisfy the constraint $y^{(i)}\left(\theta \cdot x^{(i)} + \theta_0\right) \geq 1$

posted 3 days ago by **ducanh-le**

Thanks, but in the realizable case, where the sum of the hinges losses is zero, the objective function really only does contain $\frac{1}{2}\ ||\ \theta\ ||^2$, so I don't see how that constraint can affect that its minimisation would lead to anything other than θ=0?

posted 2 days ago by **mbh038**

Minimizing $\|\theta\|^2$ subject to the constraint $y^{(i)}\left(\theta \cdot x^{(i)} + \theta_0\right) \geq 1\,(i = 1, \ldots, n)$will not lead to $\theta = 0$ because that would violate the constraint.

That being said, I disagree that in the realizable case, this is what the objective function reduces to. But that's an aside.

posted 2 days ago by **NSchandler**

Thanks

posted 2 days ago by **mbh038**

@mbh038

The program will throw away any θ which violate the constraint. Without the constraint, then $\theta = 0$. So data is filtered in the process, not before.

I think the essence of this realizable case is that we put in the constraint to **omit the if-else condition**; and then use any available computational package to solve the problem - an alternative approach to stochastic gradient descent.

posted 2 days ago by **ducanh-le**

Add a comment

Showing all responses

Add a response:

Preview

Submit

Learn About Verified Certificates