1. Objectives

Bookmark this page

Linear Classification and Generalization

At the end of this lecture, you will be able to

- understanding optimization view of learning
- apply optimization algorithms such as gradient descent, stochastic gradient descent, and quadratic program

Distance from a line to a point in terms of components

1.0/1 point (graded)

In a 2 dimensional space, a line L is given by $L : ax + by + c = 0$, and a point P is given by $P = (x_0, y_0)$. What is d, the shortest distance between L and P? Express d in terms of a, b, c, x_0, y_0.

(a*x_0 + b * y_0 + c)/sqrt(a^ **Answer:** abs(a*x_0 + b*y_0 +c) / sqrt(a^2 + b^2)

$$\frac{a \cdot x_0 + b \cdot y_0 + c}{\sqrt{a^2 + b^2}}$$

STANDARD NOTATION

Solution:

Use the projection equation. Here θ is $[a, b]$, θ_0 is c and the point is $[x_0, y_0]$.

Show answer

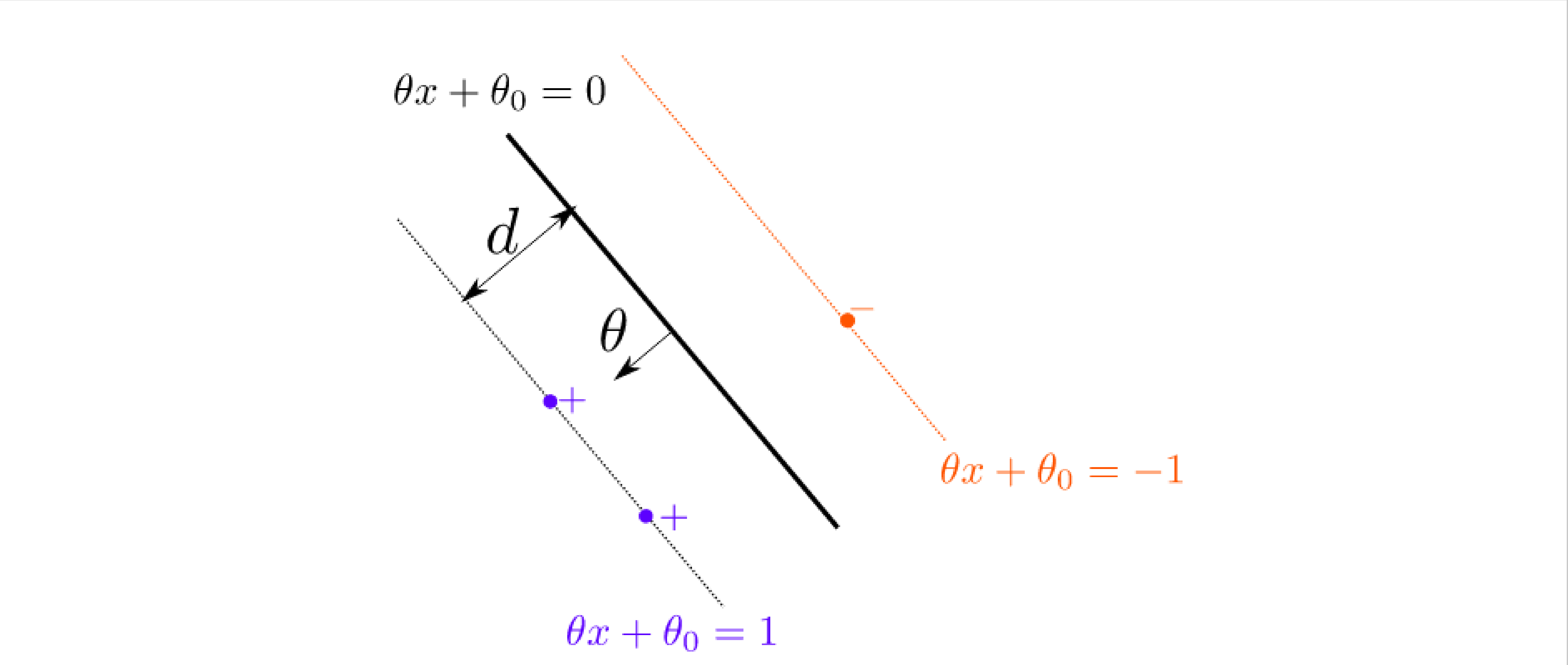
Submit You have used 1 of 3 attempts

Answers are displayed within the problem

Remember that the objective

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

In the picture below, what happens to d, the distance between the decision boundary and the margin boundary, as we increase λ?



- ○ d decreases
- ● d increases
- ○ d converges to λ

✔

Hint: You can answer with your intuition in this question. To see whether d converges to λ, think of a simple setting where we are working in 1 dimension with just two points with labels $x_1 = -1, x_2 = 2, y_1 = -1, y_2 = 1$ and assume that λ is large enough where it dominates the loss function and pushes θ close enough to 0 where all points are margin violators.

Solution:

Increasing λ means we put more weight on maximizing the margin. Thus d increases.
It is not true that d always converges to λ as λ increases. Here is a counter example:
Consider a simple setting where we are working in 1 dimension with just two points with labels $x_1 = -1, x_2 = 2, y_1 = -1, y_2 = 1$ and assume that λ is large enough where it dominates the loss function and pushes θ close enough to 0 where all points are margin violators.

$$\begin{aligned} J &= \frac{1}{2}[(1-\theta+\theta_0)+(1-2\theta-\theta_0)]+\frac{\lambda}{2}\theta^2 \\ &= \frac{2-3\theta}{2}+\frac{\lambda}{2}\theta^2. \end{aligned}$$

Solve this explicitly by taking $\frac{\partial J}{\partial \theta} = 0$:

$$\begin{aligned} \frac{-3}{2} + \lambda\theta &= 0 \\ \theta &= \frac{3}{2\lambda} \\ d &= \frac{1}{\theta} = \frac{2}{3}\lambda. \end{aligned}$$

Show answer

Submit You have used 1 of 2 attempts

Assume $\theta \in \mathbb{R}$. Our goal is to find θ that minimizes

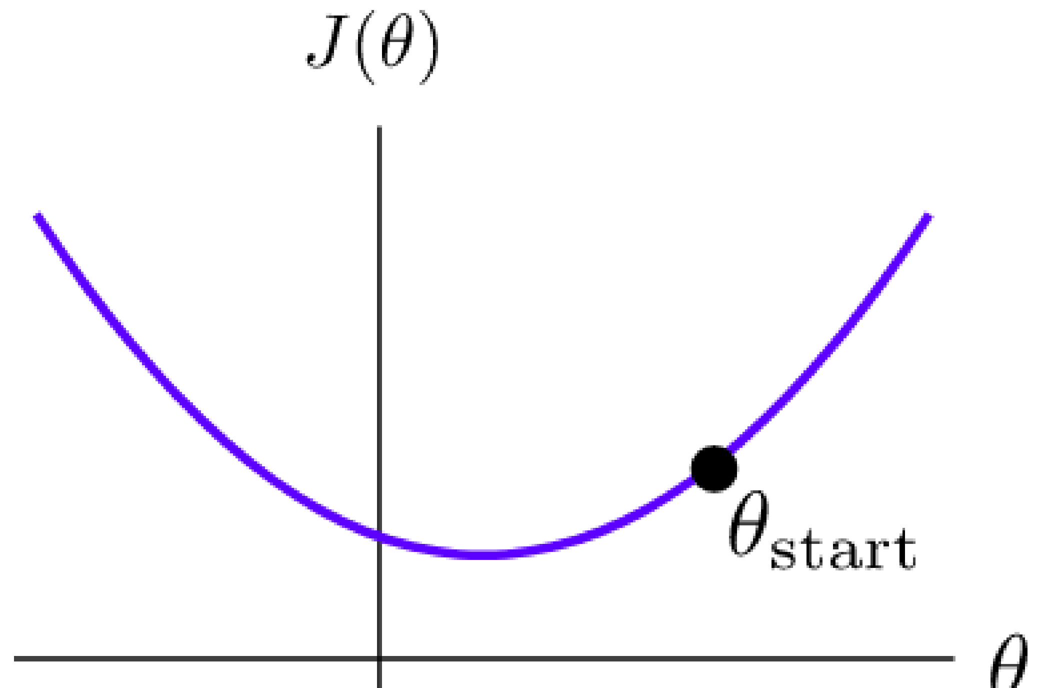
$$J(\theta, \theta_0) = \frac{1}{n} \sum_{i=1}^{n} \text{Loss}_h\left(y^{(i)}\left(\theta \cdot x^{(i)} + \theta_0\right)\right) + \frac{\lambda}{2} \mid\mid \theta \mid\mid^2$$

through gradient descent. In other words, we will

1. Start θ at an arbitrary location: $\theta \leftarrow \theta_{start}$

2. Update θ repeatedly with $\theta \leftarrow \theta - \eta \frac{\partial J(\theta,\theta_0)}{\partial \theta}$ until θ does not change significantly

In the 2 dimensional space below, we start our gradient descent at θ_{start}. What is the direction θ moves to in its first update?



- [ ] away from the origin
- [x] towards the origin
- [ ] upwards
- [ ] downwards

✔

What happens if we increase the stepsize η?

- [x] the magnitude of change in each update gets larger
- [ ] the magnitude of change in each update gets smaller

✔

Solution:

Gradient descent makes θ move to opposite direction of the gradient. Thus it will move towards the origin at θ_{start}. Also, increasing the stepsize makes the update happen in greater magnitude.

As we saw in the lecture above,

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

With stochastic gradient descent, we choose $i \in \{1, \ldots, n\}$ at random and update θ such that

$$\theta \leftarrow \theta - \eta \nabla_\theta \left[\text{Loss}_h\left(y^{(i)}\left(\theta \cdot x^{(i)} + \theta_0\right)\right) + \frac{\lambda}{2} \mid\mid \theta \mid\mid^2\right]$$

What is $\nabla_\theta \left[\text{Loss}_h\left(y^{(i)}\left(\theta \cdot x^{(i)} + \theta_0\right)\right)\right]$ if $\text{Loss}_h\left(y^{(i)}\left(\theta \cdot x^{(i)} + \theta_0\right)\right) > 0$?

- ○ $y^{(i)}x^{(i)}$
- ◉ $-y^{(i)}x^{(i)}$
- ○ 0
- ○ $\lambda\theta$
- ○ $-\lambda\theta$

Solution:

If $\text{Loss}_h\left(y^{(i)}\left(\theta \cdot x^{(i)} + \theta_0\right)\right) > 0$,

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

. Thus

$$\nabla_\theta \text{Loss}_h\left(y^{(i)}\left(\theta \cdot x^{(i)} + \theta_0\right)\right) = -y^{(i)} x^{(i)}$$

.

Show answer

Submit You have used 2 of 3 attempts

Answers are displayed within the problem

Comparison with Perceptron

1/1 point (graded)

Observing the update step of SGD,

$$\theta \leftarrow \theta - \eta \nabla_\theta \left[\text{Loss}_h \left(y^{(i)} \left(\theta \cdot x^{(i)} + \theta_0 \right) \right) + \frac{\lambda}{2} \| \theta \|^2 \right]$$

Which of the following is true?

- [ ] As in perceptron, θ is not updated when there is no mistake
- [x] Differently from perceptron, θ is updated even when there is no mistake

✔

Solution:

We can see from

$$\theta \leftarrow \begin{cases} (1-\lambda\eta)\,\theta \text{ if Loss}=0 \\ (1-\lambda\eta)\,\theta + \eta y^{(i)} x^{(i)} \text{ if Loss}>0 \end{cases}$$

that θ is updated even when the sum of losses is 0. This is different from perceptron.

Show answer

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 $i = 1, \ldots n$.
- (x) 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.

Show answer

Submit You have used 2 of 2 attempts

Answers are displayed within the problem

The realizable case 2

Remember the objective function

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

In the realizable case, we can always find $\left(\theta, \theta_0\right)$ 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 $\frac{1}{2} \mid\mid \theta \mid\mid^2$

Show answer

Support Vectors

0/1 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
- [x] 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
- [x] 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 $J(\theta, \theta_0)$. 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.

Show answer

Submit You have used 2 of 2 attempts