STAT 535

11/14/23

Lecture 14

SGD
Heavy Ball
Coordinate descent

SVM

LV SVM posted
LV.2 RFF

Lecture IV: Training predictors, Part II

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

November, 2023

Stochastic gradient methods

Examples: Linear classification with hinge loss, P~~erceptron~~

Accelerated gradient

No gradient methods: Coordinate descent

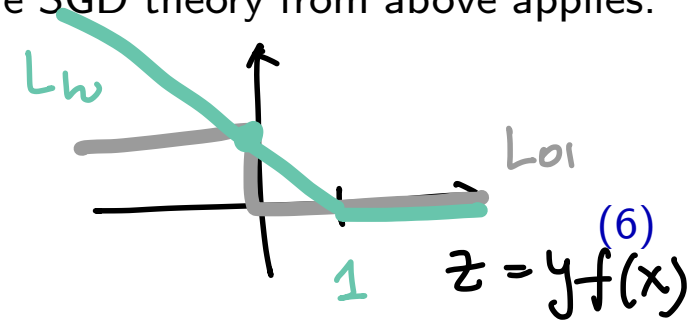
Stopping descent algorithms

Reading HTF Ch.: –, Murphy Ch.: 8.5.2-3 Stochastic gradient descent, Bach Ch.: For more advanced treatment Nocedal and Wrigth.

Example: Linear classification with hinge loss

[ML framework, minimize J or $\hat{L}$ over parameters, (x, y) =data, f = predictor]

The following example is a classic in statistical learning. We will examine it in two formulation. The first is an example of a problem where λ is known, and the SGD theory from above applies.



Problem setting

- $y \in \{\pm 1\}$ (binary classification)
- We fit the linear classifier

$$f(x) = w^T x \tag{6}$$

- Loss function = **hinge loss**

$$L_h(y, f(x)) = \begin{cases} 0 & \text{if } yf(x) \geq 1 \\ 1 - yf(x) & \text{if } yf(x) < 1 \end{cases} = [yf(x) - 1]_- \tag{7}$$

Define **margin** of example x

$$z = yf(x) \tag{8}$$

Under L_h an error is penalized linearly by how far $f(x)$ is in the "wrong direction" to which we add a penalty even for correctly classified examples if the margin $yf(x)$ is below 1.

- using the hinge loss.
- Simplifying assumption (for now, will remove it when we study SVM): the data $\mathcal{D} = \{(x^i, y^i)\}_{i=1:n}$ are **linearly separable**, i.e. there exists a w^* that classifies the sample with no error. Note that in general this w^* is not unique.

Linear Support Vector Machine SGD training algorithm

The optimization problem is a regularized one:

$$J(w) = \frac{1}{n}\sum_i L_h(y^i, w^T x^i) + \frac{\lambda}{2}||w||^2 \quad (9)$$

with $\lambda > 0$ a regularization parameter chosen by the user. Note that the non-quadratic loss term is linear (with unknown slope at w^*) and therefore the function $J(w)$ is by definition λ-strongly convex.

The stochastic part of the gradient is

$$\frac{\partial L_h}{\partial w} = \begin{cases} -y^i x^i & \text{if } i \text{ ``error''} \\ 0 & \text{if } i \text{ ``correct''} \end{cases} \quad (10)$$

where "correct" means that $y^i f(x^i) > 1$.

SGD for Linear SVM

Initialize with $w^0 = 0$, $\bar{w} = 0$

Iterate for $k = 1, 2, \ldots K$

1. Pick the next i in $1 : n$
2. compute direction $d^k = \lambda w^k - 1_{[i \text{ ``error''}]} y^i x^i$
3. update

$$w^{k+1} = w^k - \frac{c}{\lambda k}\left(\lambda w^k - 1_{[i \text{ ``error''}]} y^i x^i\right) = w^k(1 - c/k) + \frac{c}{\lambda k} y^i x^i 1_{[i \text{ ``error''}]} \quad (11)$$

4. average $\bar{w} \leftarrow \bar{w} + w^{k+1}$

Output $\bar{w}/K$

SGD for streaming data

$(x^i, y^i) \sim$ iid P_{XY} → Alg uses (x^i, y^i) for training then forgets it.

No data is stored.

SGD = the same !! (or faster) $(n' = 1)$

⇒ Natural on streaming data

Streaming SGD ↓ expected

$E[d] = \nabla L$

Var(d) also bded

$\propto \sigma^2$

Original SGD ↑

$E[d] = \nabla L$

$Var(d) \leq \sigma^2$ bounded

Accelerated gradient: the "heavy ball" method

$$x^{k+1} = x^k - \eta^k d^k + \gamma^k (x^k - x^{k-1}) \tag{14}$$

- Applies to both standard and stochastic gradient methods, i.e.

$$d^k = \begin{cases} \nabla f(x^k) & \text{gradient descent} \\ \text{noisy gradient} & \text{SGD} \\ \nabla f(x^k + \delta^k (x^k - x^{k-1})) & \text{extragradient methods} \end{cases} \tag{15}$$

- Setting the parameters
  - In the extragradient[3] methods, $\eta^k, \delta^k, \gamma^k$ are obtained by search (or knowledge about M, m)
  - For other methods fix $\gamma^k = \gamma \in (0.5, 1]$ OR use smaller γ early in the training and increase it to near 1 when the steps become smaller.
- More intuition
  - for ill conditioned problems $M \ll m$, the heavy ball "accumulates" the components of the step in the correct direction
  - for SGD, the heavy ball approximates the exact gradient

[3] Nesterov's "optimal", FISTA

Heavy Ball

$$x^{k+1} = x^k - \eta^k d^k + \gamma^k (x^k - x^{k-1})$$

x = parameter vector !!

$$x^{k+1} - x^k = -\eta^k d^k + \gamma^k (x^k - x^{k-1})$$

η^k: step size; d^k: current direction; $\gamma^k (x^k - x^{k-1})$: momentum term

$0 < \gamma^k < 1$ decay parameter

$\gamma^k = \gamma$ constant

$\eta^k = \eta$ constant

$$x^{k+1} - x^k = -\eta^k d^k + \gamma (-\eta^{k-1} d^{k-1} + \gamma (x^{k-1} - x^{k-2}))$$
$$= \dots$$
$$= -\eta (d^k + \gamma d^{k-1} + \gamma^2 d^{k-2} + \dots)$$

$\tilde{d}^k$ = sum

$$1 + \gamma + \gamma^2 + \dots = \frac{1}{1-\gamma} > 1$$

$\gamma \lesssim 1 \Rightarrow x^{k+1} - x^k \propto$ weighted average of recent $d^k, d^{k-1}, \dots$

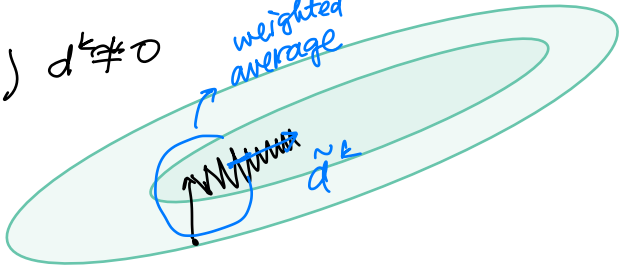
Effect

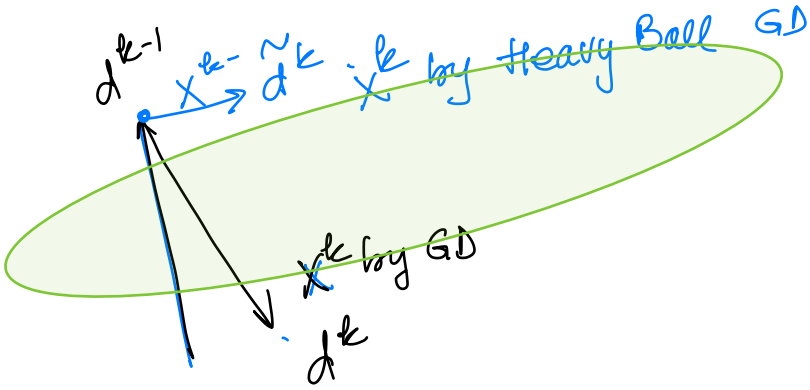
- near x^*

  GD $d^k \approx 0 \Rightarrow \tilde{d}^k \approx 0$

  SGD $d^k \not\approx 0$, $E[d^k] \approx 0$ $\Rightarrow \|\tilde{d}^k\| < \|d^k\|$ closer to 0 !

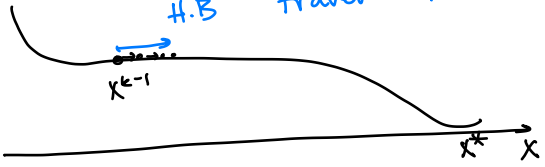
- far from x^* "valley" (GD) $d^k \neq 0$



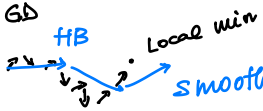


- Plateau $d^k \approx 0$ and far from x^*

a)



b) • Uneven Plateau

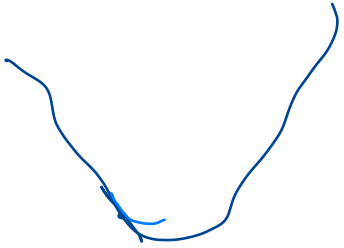


smooths trajectory
escapes small local min

3. Why successful (theory)

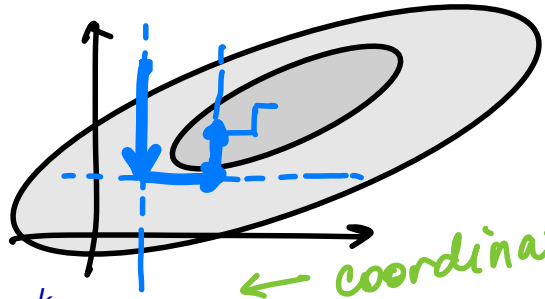
HB → more information than GD !!

⇒ faster convergence



GD only slope ∇ Newton ∇, ∇^2

Coordinate descent



- d^k is always one of the coordinate axes u_{ik}. Hence $x^{k+1} = x^k + \eta_k u_{ik}$.
- Note that line search is necessary, and that the minimum can be on either side of x^k so η_k can take negative values.

Convergence Theoretical and empirical results suggest that coordinate descent has similar convergence rate as the steepest descent (i.e linear in the best case).
While in a general case coordinate descent is suboptimal, there are several situations when it is worth considering along 1 coordinate

1. When line minimization can be done analytically. This can save one the often expensive gradient computation.
2. When the coordinate axes affect the function value approximately independently, or (in statistics) when the coordinate axes are uncorrelated. Then minimizing along each axis separately is (nearly) optimal.
3. When there exists a natural grouping of the variables. Then one can optimize one group of variables while keeping the other constant. Again, we hope that the groups are "independent", or that optimizing one group at a time can be done analytically, or it's much easier than computing the gradient w.r.t all variables simultaneously. This idea is the basis of many *alternate minimization* methods, including the well known EM algorithm.

AM MM

Ex Max likelihood with hidden variables

$L(\theta, z)$

Advanced Topics

Lecture V: Support Vector Machines

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

November, 2023

Linear SVM's
- The margin and the expected classification error
- Maximum Margin Linear classifiers
- Linear classifiers for non-linearly separable data

Non linear SVM
- The "kernel trick"
- Kernels
- Prediction with SVM

Extensions
- L_1 SVM
- Multi-class and One class SVM
- SV Regression

Reading HTF Ch.: Ch. 12.1–3, Murphy Ch.: Ch 14 (14.1,14.2–14.2.4 kernels, 14.4 and equations (14.28,14.29) kernel trick, 14.5.1.–3 Support Vector Machines), Bach Ch.: 7.1–7.4, 7.7
Additional Reading: C. Burges - "A tutorial on SVM for pattern recognition"
These notes: Appendices (convex optimization) are optional.

The margin and the expected classification error

THEORY

The following two theorems suggest that large margin is a predictor of good generalization error.

Theorem Let $\mathcal{F}_\Delta$ be the class of hyperplanes $f(x) = w^T x$, x, $w \in \mathbb{R}^n$, that are Δ away from any data point[1] in the training set $\mathcal{D}$. Then,

$$VCdim\, \mathcal{F}_\Delta \leq 1 + \min\left(d, \frac{R^2_{\mathcal{D}}}{\Delta^2}\right) \tag{2}$$

Δ geometric

where $R_{\mathcal{D}}$ is the radius of the smallest ball that encloses the dataset.

Theorem Let $\mathcal{F} = \{\text{sgn}\,(w^T x),\ ||w|| \leq \Lambda,\ ||x|| \leq R\}$ and let $\rho > 0$ be any "margin". Then for any $f \in \mathcal{F}$, w.p $1 - \delta$ over training sets

expected empirical

$$L_{01}(f) \leq \hat{L}_\rho + \sqrt{\frac{c}{n}\left(\frac{R^2\Lambda^2}{\rho^2}\ln n^2 + \ln\frac{1}{\delta}\right)} \tag{3}$$

where c is a universal constant and $\hat{L}_\rho$ is the fraction of the training examples for which

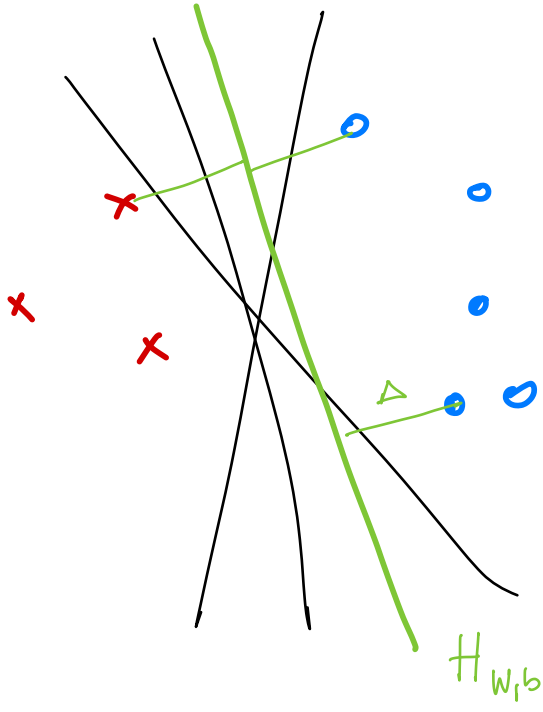
$$y^i w^T x_i < \rho \tag{4}$$

$y^i f(x^i)$ margin of i

margin error

A data point i that satisfies (4) for some ρ is called a **margin error**. For $\rho = 0$ the margin error rate $\hat{L}_\rho$ is equal to $\hat{L}_{01}$.

[1] In other words, a set $\mathcal{D}$ is shattered only if all the linear classifiers pass at least Δ away from its points.



Assume data linearly separable

$\Delta_{w,b} = \min_i \text{dist}(x^i, H_{w,b})$

$f(x) = w^T x + b$ linear classifier

$H_{w,b} = \{x, f(x) = 0\}$

SVM pb max margin

want $\arg\max_{w,b} \Delta_{w,b}$ s.t.

(P1) $y^i(w^T x^i + b) > 0$

classify correctly

$y^i(w^T x^i + b) \geq 1$

Remark 3. $\min_i y^i(w^T x^i + b) = 1 \Rightarrow$

(P3) $\max_{w,b} \min_i \frac{|w^T x^i + b|}{\|w\|}$ Δ

$\min_i |w^T x^i + b| = 1$

Remark 1.

(P2) .... s.t. $y^i(w^T x^i + b) \geq 1$

Remark 2.

$d(x, H_{w,b}) = \frac{|w^T x + b|}{\|w\|}$

(P4) $\max_{w,b} \frac{1}{\|w\|} \quad s.t \quad y^i(w^T x^i + b) \geq 1$

(P5) $\min_{w,b} \|w\|^2 \quad s.t \quad$ —"— CONVEX

(P6) $\min \|w\|^2 \quad s.t \quad$ —"— QUADRATIC

easier to analyse!

Maximum Margin Linear classifiers

Support Vector Machines appeared from the convergence of Three Good Ideas

Assume (for the moment) that the data are linearly separable.

- Then, there are an infinity of linear classifiers that have $\hat{L}_{01} = 0$. Which one to choose?

1st idea Select the classifier that has **maximum margin** Δ on the training set.

By SRM, we should choose the (w, b) parameters that minimize $\hat{L}(w, b) + R(h_{w,b})$, where $h_{w,b}$ is given by (2):

- For any parameters (w, b) that perfectly classify the data $\hat{L}(w, b) = 0$.
- Among these, the best (w, b) is the one that minimizes $R(h_{w,b})$
- Among these, the best (w, b) is the one that ~~minimizes~~ maxi Δ in **??**
- Hence, we should choose

$$\underset{\Delta,w,b:\hat{L}(w,b)=0}{\operatorname{argmax}} \Delta, \quad \text{s.t. } d(x, H_{w,b}) \geq \Delta \text{ for } i = 1 : n, \tag{7}$$

where $d()$ denotes the Euclidean distance and $H_{w,b} = \{x \mid w^T x + b = 0\}$ is the decision boundary of the linear classifier.

- Because $d(x, H_{w,b}) = \frac{|w^T x + b|}{||w||}$ (proof in a few slides) (7) becomes

$$\underset{\Delta,w,b:\hat{L}(w,b)=0}{\operatorname{argmax}} \Delta, \quad \text{s.t. } \frac{|w^T x^i + b|}{||w||} \geq \Delta \text{ for } i = 1 : n, \tag{8}$$

Maximum Margin Linear classifiers

We continue to transform (8)

- If all data correctly classified, then $y^i(w^T x^i + b) = |w^T x^i + b|$. Therefore (8) has the same solution as

$$\underset{\Delta, w, b}{\operatorname{argmax}} \Delta, \quad \text{s.t. } \frac{y^i(w^T x^i + b)}{||w||} \geq \Delta \text{ for } i = 1 : n, \tag{9}$$

- Note now that the problem (9) is underdetermined. Setting $w \leftarrow Cw, b \leftarrow Cb$ with $C > 0$ does not change anything.
- We add a cleverly chosen constraint to remove the indeterminacy; this is $||w|| = 1/\Delta$, which allows us to eliminate the variable Δ. We get

$$\underset{w, b}{\operatorname{argmax}} \frac{1}{||w||}, \quad \text{s.t. } y^i(w^T x^i + b) \geq 1 \text{ for } i = 1 : n, \tag{10}$$

Note: the successive problems (7),(8),(9),. . . are equivalent in the sense that their optimal solution is the same.

Alternative derivation of (10)

st idea Select the classifier that has maximum margin on the training set, by the alternative definition of margin.
Formally, define $\min_{i=1:n} y^i f(x^i)$ be the **margin of classifier f on $\mathcal{D}$**. Let $f(x) = w^T x + b$, and choose w, b that

$$\text{maximize}_{w \in \mathbb{R}^n, b \in \mathbb{R}} \min_{i=1:n} y^i (w^T x^i + b) \; s.t. \; \hat{L}(w, b) = 0$$

- Remarks
  - (if data is linearly separable), there exist classifiers with margins > 0
  - one can arbitrarily increase the margin of such a classifier by multiplying w and b by a positive constant.
  - Hence, we need to "normalize" the set of candidate classifiers by requiring instead

$$\text{maximize} \min_{i=1:n} d(x, H_{w,b}), \text{ s.t. } y^i (w^T x^i + b) \geq 1 \text{ for } i = 1:n, \tag{11}$$

    where $d()$ denotes the Euclidean distance and $H_{w,b} = \{ x \mid w^T x + b = 0 \}$ is the decision boundary of the linear classifier.
  - Under the conditions of (11), because there are points for which $|w^T x + b| = 1$, maximizing $d(x, H_{w,b})$ over w, b for such a point is the same as

$$\max_{w,b} \frac{1}{||w||}, \text{ s.t. } \min_i y_i (w^T x + b) = 1 \tag{12}$$

Second idea

The **Second idea** is to formulate (10) as a **quadratic** optimization problem.

$$\min_{w,b} \frac{1}{2}||w||^2 \text{ s.t } y^i(w^T x^i + b) \geq 1 \text{ for all } i = 1 : n \tag{13}$$

This is the **Linear SVM (primal) optimization problem**

- This problem has a strongly convex **objective** $||w||^2$, and **constraints** $y^i(w^T x^i + b)$ linear in (w, b).
- Hence this is a convex problem, and can be studied with the tools of convex optimization.

The distance of a point x to a hyperplane $H_{w,b}$

$$d(x, H_{w,b}) = \frac{|w^T x + b|}{||w||} \tag{14}$$

Intuition: denote

$$\tilde{w} = \frac{w}{||w||}, \ \tilde{b} = \frac{b}{||w||}, \ x' = \tilde{w}^T x. \tag{15}$$

Obviously $H_{w,b} = H_{\tilde{w},\tilde{b}}$, and x' is the length of the projection of point x on the direction of w.

The distance is measured along the normal through x to H; note that if $x' = -\tilde{b}$ then $x \in H_{w,b}$ and $d(x, H_{w,b}) = 0$; in general, the distance along this line will be $|x' - (-\tilde{b})|$.