Usual setup in machine learning

Choose a model (w) by minimizing a loss function $(L(w))$ that depends on the data.

- Linear regression:

$$L(w) = \sum_i \left(y^{(i)} - \left(w \cdot x^{(i)}\right)\right)^2 \tag{1}$$

- Logistic regression:

$$L(w) = \sum_i \ln\left(1 + e^{-y^{(i)}\left(w \cdot x^{(i)}\right)}\right) \tag{2}$$

Default way to solve this minimization

Local search.

- Initialize (w) arbitrarily
- Repeat until (w) converges:
  - Find some (w') close to (w) with $(L(w') < L(w))$.
  - Move (w) to (w').

A Good Situation for Local Search

When the loss function is convex.

Idea

Pick search direction by looking at derivative of $(L(w))$.

Multivariate Differentiation: Example

$$F\left(w_1, w_2, w_3\right) = 3w_1w_2 + w_3 \tag{3}$$

Suppose we are learning a model with (k) parameters $(w = (w_1, \ldots, w_k))$.

- Define a loss function $(L(w))$
- Then $(L : \mathbb{R}^k \to \mathbb{R})$

The derivative $(\nabla F(w))$, at any (w), is a vector in $(\mathbb{R}^k)$.

Gradient Descent

For minimizing a function ($L(w)$):

- ($w_0 = 0, t = 0$)
- while ($\nabla L(w_t) \approx 0$):
  - ($w_{t+1} = w_t - \eta_t \nabla L(w_t)$)
  - ($t = t + 1$)

Here (η_t) is the step size at time (t).

Gradient Descent for Logistic Regression

For ($\left(x^{(1)}, y^{(1)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right) \in \mathbb{R}^d \times \{-1, 1\}$), loss function:

$$L(w) = \sum_{i=1}^{n} \ln\left(1 + e^{-y^{(i)}\left(w \cdot x^{(i)}\right)}\right) \tag{4}$$

Gradient descent procedure:

- Set ($w_0 = 0$)
- For ($t = 0, 1, 2, \ldots$), until convergence:

$$w_{t+1} = w_t + \eta_t \sum_{i=1}^{n} y^{(i)} x^{(i)} \underbrace{p_{r_{w_t}}\left(-y^{(i)} \mid x^{(i)}\right)}_{\mathrm{doubt}_t\left(x^{(i)}, y^{(i)}\right)} \tag{5}$$

How to set step size (η_t)?

A Variant of Gradient Descent

Gradient descent for logistic regression, given ($\left(x^{(1)}, y^{(1)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)$):

- Set ($w_0 = 0$)
- For ($t = 0, 1, 2, \ldots$), until convergence:

$$w_{t+1} = w_t + \eta_t \sum_{i=1}^{n} y^{(i)} x^{(i)} \Pr_{w_t}\left(-y^{(i)} \mid x^{(i)}\right) \tag{6}$$

Each update involves the entire data set, which is inconvenient.
Stochastic gradient descent: update based on just one point:

- Get next data point ($x; y$) by cycling through data set
- ($w_{t+1} = w_t + \eta_t y \times \Pr_{w_t}(-y \mid x)$)

Decomposable Loss Functions

Loss function for logistic regression:

$$L(w) = \sum_{i=1}^{n} \ln\left(1 + e^{-y^{(i)}\left(w\cdot x^{(i)}\right)}\right) = \sum_{i=1}^{n}\left(\text{loss of } w \text{ on } \left(x^{(i)}, y^{(i)}\right)\right) \tag{7}$$

Most ML loss functions are like this: Given $(\left(x^{(1)}, y^{(1)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right))$,

$$L(w) = \sum_{i=1}^{n} \ell\left(w; x^{(i)}, y^{(i)}\right) \tag{8}$$

where $(\ell(w; x, y))$ captures the loss on a single point.

Gradient Descent and Stochastic Gradient Descent

For minimizing

$$L(w) = \sum_{i=1}^{n} \ell\left(w; x^{(i)}, y^{(i)}\right) \tag{9}$$

Gradient descent:

- $(w_0 = 0)$
- While not converged:

$$w_{t+1} = w_t - \eta_t \sum_{i=1}^{n} \nabla\ell\left(w_t; x^{(i)}, y^{(i)}\right) \tag{10}$$

Stochastic gradient descent:

- $(w_0 = 0)$
- Keep cycling through data points (x, y):

$$w_{t+1} = w_t - \eta_t \nabla\ell\left(w_t; x, y\right) \tag{11}$$

Mini-batch gradient descent:

- $(w_0 = 0)$
- Repeat:
  - Get the next batch of points (B)
  - $(w_{t+1} = w_t - \eta_t \sum_{(x,y)\in B} \nabla\ell\left(w_t; x, y\right))$

Is our Loss Function Convex?

Convexity

A function ($f : \mathbb{R}^d \to \mathbb{R}$) is convex if for all ($a, b \in \mathbb{R}^d$) and ($0 < \theta < 1$),

$$f(\theta a + (1-\theta)b) \leq \theta f(a) + (1-\theta) f(b) \tag{12}$$

It is strictly convex if strict inequality holds for all ($a \neq b$).
(f) is concave $\Leftrightarrow -f$ is convex

Checking Convexity for Functions of One Variable

A function ($f : \mathbb{R} \to \mathbb{R}$) is convex if its second derivative is (≥ 0) everywhere.
Example: ($f(z) = z^2$)

Function of one variable

$$F : \mathbb{R} \to \mathbb{R} \tag{13}$$

- Value: number
- Derivative: number
- Second derivative: number

Convex if second derivative is always (≥ 0)

Function of (d) variables

$$F : \mathbb{R}^d \to \mathbb{R} \tag{14}$$

- Value: number
- Derivative: (d)-dimensional vector
- Second derivative: ($d \times d$) matrix

Convex if second derivative matrix is always positive semidefinite

First & Second Derivatives of Multivariate Functions

For a function ($f : \mathbb{R}^d \to \mathbb{R}$),

- the first derivative is a vector with (d) entries:

$$\nabla f(z) = \begin{pmatrix} \frac{\partial f}{\partial z_1} \\ \vdots \\ \frac{\partial f}{\partial z_d} \end{pmatrix} \tag{15}$$

- the second derivative is a $(d \times d)$ matrix, the Hessian $(H(z))$:

$$H_{jk} = \frac{\partial^2 f}{\partial z_j \partial z_k} \tag{16}$$

Example

$(w \in \mathbb{R}^d)$ and $(F(w) = \|w\|^2)$. Find the derivative.

Example

Find the second derivative matrix of $(F(w) = \|w\|^2)$.

Gradient Descent

For minimizing a function $(L(w))$:

- $(w_0 = 0, t = 0)$
- while $(\nabla L(w_t) \approx 0)$:
  - $(w_{t+1} = w_t - \eta_t \nabla L(w_t))$
  - $(t = t + 1)$

Here (η_t) is the step size at time (t).

Gradient Descent: Rationale

"Differentiable" $\Rightarrow$ "locally linear".

For small displacements $(u \in \mathbb{R}^d)$,

$$L(w + u) \approx L(w) + u \cdot \nabla L(w) \tag{17}$$

Therefore, if $(u = -\eta \nabla L(w))$ is small,

$$L(w + u) \approx L(w) - \eta \|\nabla L(w)\|^2 < L(w) \tag{18}$$

The Step Size Matters

Gradient Descent Update: $(w_{t+1} = w_t - \eta_t \nabla L(w_t))$.

- Step size (η_t) too small: not much progress
- Too large: overshoot the mark

One option: pick (η_t) using a line search

$$\eta_t = \arg\min_{\alpha > 0} L\left(w_t - \alpha \nabla L(w_t)\right) \tag{19}$$

Example: Logistic Regression

For $\left(\left(x^{(1)}, y^{(1)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right) \in \mathbb{R}^d \times \{-1, 1\}\right)$, loss function:

$$L(w) = \sum_{i=1}^{n} \ln\left(1 + e^{-y^{(i)}\left(w \cdot x^{(i)}\right)}\right) \tag{20}$$

What is the derivative?

- Set $(w_0 = 0)$
- For $(t = 0, 1, 2, \ldots)$, until convergence:

$$w_{t+1} = w_t + \eta_t \sum_{i=1}^{n} y^{(i)} x^{(i)} \underbrace{Pr_{w_t}\left(-y^{(i)} \mid x^{(i)}\right)}_{\text{doubt}_t\left(x^{(i)}, y^{(i)}\right)} \tag{21}$$

Recall

Every square matrix (M) encodes a quadratic function:

$$x \longmapsto x^T M x = \sum_{i=1}^{d} M_{ij} x_i x_j \tag{22}$$

(M) is a $(d \times d)$ matrix and (x) is a vector in $(\mathbb{R}^d)$.

Positive Semidefinite Matrices

A symmetric matrix (M) is positive semidefinite (psd) if:

$$x^T M x \geq 0 \text{ for all vectors } x \tag{23}$$

When is a diagonal matrix PSD?

If (M) is PSD, must (cM) be PSD for a constant (c)?

If (M, N) are of the same size and PSD, must $(M + N)$ be PSD?

Checking if a Matrix is PSD

A matrix (M) is PSD if and only if it can be written as $(M = UU^T)$ for some matrix (U).
Quick check: say $(U \in \mathbb{R}^{r \times d})$ and $(M = UU^T)$.

1. (M) is square.
2. (M) is symmetric.

3. Pick any $(x \in \mathbb{R}^r)$. Then

$$\begin{aligned} x^T M x &= x^T U U^T x \\ &= \left(x^T U\right)\left(U^T x\right) \\ &= \left(U^T x\right)^T \left(U^T x\right) \\ &= \left\|U^T x\right\|^2 \geq 0 \end{aligned} \tag{24}$$

Another useful fact

Any covariance matrix is PSD.

A Hierarchy of Square Matrices

Square

$$M \in \mathbb{R}^{d \times d} \tag{25}$$

Positive Semidefinite

$$x^T M x \geq 0 \text{ for all } x \in \mathbb{R}^d \tag{26}$$

1 Checking Convexity

1.1 Function of one variable

$F : \mathbb{R} \longrightarrow \mathbb{R}$

- Value: number
- Derivative: number
- Second derivative: number

Convex if second derivative is always ≥ 0

1.2 Function of d variables

$F : \mathbb{R}^d \longrightarrow \mathbb{R}$

- Value: number
- Derivative: d-dimensional vector
- Second derivative: $d \times d$ matrix

Convex if second derivative matrix is always positive semidefinite

2 Second-Derivative Test for Convexity

A function of several variables, $F(z)$, is convex if its second-derivative matrix $H(z)$ is positive semidefinite for all z.

2.1 More formally:

Suppose that for $f : \mathbb{R}^d \rightarrow \mathbb{R}$ the second partial derivatives exist everywhere and are continuous functions of z.

Then:

1. $H(z)$ is a symmetric matrix
2. f is convex $\Leftrightarrow$ $H(z)$ is positive semidefinite for all $z \in \mathbb{R}^d$

3 Example

Is $f(x) = ||x||^2$ convex?

4 Example

Fix any vector $u \in \mathbb{R}^d$ Is this function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ convex?

$f(z) = (u \cdot z)^2$

5 Least-Squares Regression

Recall loss function: for data points $(x^{(i)}, y^{(i)}) \in \mathbb{R}^d \times \mathbb{R}$.

$L(w) = \sum_{i=1}^{n} (y^{(i)} - (w \cdot x^{(i)}))^2$

6 Minimizing a loss function

Usual setup in machine learning: choose a model w by minimizing a loss function $L(w)$ that depends on the data. [cite: 2, 3]

- Linear regression: $L(w) = \sum_i (y^{(i)} - (w \cdot x^{(i)}))^2$ [cite: 3]
- Logistic regression: $l(w) = \sum_i ln(1 + e^{-y^{(i)}(w \cdot x^{(i)})})$ [cite: 3]

Default way to solve this minimization: local search. [cite: 3, 4]

7 Local search

- Initialize w arbitrarily
- Repeat until w converges:
  - Find some w' close to w with $L(w') < L(w)$. [cite: 4, 5]
  - Move w to w'. [cite: 5]

8 A good situation for local search

When the loss function is convex:
Idea for picking search direction: Look at the derivative of $L(w)$ at the current point w. [cite: 6]

9 Gradient descent

For minimizing a function $L(w)$:
$w_o = 0$ $t = 0$

- while $\nabla L(w_t) \approx 0$:
  - $w_{t+1} = w_t - \eta_t \nabla L(w_t)$ [cite: 7]
  - $t = t + 1$

Here η_t is the step size at time t. [cite: 7]

10 Multivariate differentiation

Example: $w \in \mathbb{R}^3$ and $F(w) = 3w_1w_2 + w_3$. [cite: 8]
Example: $w \in \mathbb{R}^d$ and $F(w) = w \cdot x$. [cite: 8]

11 Gradient descent: rationale

"Differentiable" $\approx$ "locally linear". [cite: 9]
For small displacements $u \in \mathbb{R}^d$
$L(w + u) \approx L(w) + u \cdot \nabla L(w)$ [cite: 9]
Therefore, if $u = -\eta \nabla L(w)$ is small,
$L(w + u) \approx L(w) - \eta ||\nabla L(w)||^2 < L(w)$ [cite: 9]

12 The step size matters

Update rule: $w_{t+1} = w_t - \eta_t \nabla L(w_t)$

- Step size η_t too small: not much progress
- Too large: overshoot the mark

Some choices:

- Set η_t according to a fixed schedule, like $1/t$
- Choose by line search to minimize $L(w_{t+1})$

13 Example: logistic regression

For $(x^{(1)}, y^{(1)}), ..., (x^{(n)}, y^{(n)}) \in \mathbb{R}^d \times \{-1, +1\}$, loss function
$L(w) = \sum_{i=1}^{n} ln(1 + e^{-y^{(i)}(w \cdot x^{(i)})})$ [cite: 10]
What is the derivative?

14 Gradient descent for logistic regression

- Set $w_0 = 0$
- For $t = 0, 1, 2, ...,$ until convergence:
  $w_{t+1} = w_t + \eta_t \sum_{i=1}^{n} y^{(i)} x^{(i)} Pr_{w_t}(-y^{(i)}|x^{(i)})$ [cite: 12]

15 Gradient descent for large data sets?

- Set $w_0 = 0$
- For $t = 0, 1, 2, ...,$ until convergence:
  $w_{t+1} = w_t + \eta_t \sum_{i=1}^{n} y^{(i)} x^{(i)} Pr_{w_t}(-y^{(i)}|x^{(i)})$ [cite: 13]

Each update involves the entire data set, which is inconvenient. [cite: 13, 14]

16 Stochastic gradient descent: update based on just one point:

- Get next data point (x, y) by cycling through data set
- $w_{t+1} = w_t + \eta_t y \, x Pr_{w_t}(-y|x)$ [cite: 14]

17 Decomposable loss functions

Loss function for logistic regression:

$L(w) = \sum_{i=1}^{n} ln(1 + e^{-y^{(i)}(w \cdot x^{(i)})}) = \sum_{i=1}^{n}($ loss of won $(x^{(i)}, y^{(i)}))$ [cite: 15]

Most ML loss functions are like this: for data $(x^{(1)}, y^{(1)}), ..., (x^{(n)}, y^{(n)})$,

$l(w) = \sum_{i=1}^{n} l(w; x^{(i)}, y^{(i)})$ [cite: 15]

where $l(w; x, y)$ captures the loss on a single point. [cite: 15, 16]

18 Gradient descent and stochastic gradient descent

For minimizing

$L(w) = \sum_{i=1}^{n} l(w; x^{(i)}, y^{(i)})$

Gradient descent:

$w_o = 0$

- while not converged:

  $w_{t+1} = w_t - \eta_t \sum_{i=1}^{n} \nabla l(w_t; x^{(i)}, y^{(i)})$ [cite: 16]

Stochastic gradient descent: $w_o = 0$

- Keep cycling through data points (x, y):

  $w_{t+1} = w_t - \eta_t \nabla l(w_t; x, y)$ [cite: 16]

19 Variant: mini-batch stochastic gradient descent

Stochastic gradient descent:

$w_o = 0$

- Keep cycling through data points (x, y) $w_{t+1} = w_t - \eta_t \nabla l(w_t; x, y)$ [cite: 17]

Mini-batch stochastic gradient descent:

$w_o = 0$

- Repeat:
- Get the next batch of points B
- $w_{t+1} = w_t - \eta_t \sum_{(x,y) \in B} \nabla l(w_t; x, y)$ [cite: 17]

20 Convexity

Is our loss function convex?

21 Convexity

A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if for all a, $b \in \mathbb{R}^d$ and $0 < \theta < 1$

$f(\theta a + (1-\theta)b) \leq \theta f(a) + (1-\theta)f(b)$. [cite: 47]

It is strictly convex if strict inequality holds for all $a \neq b$. [cite: 48]

f is concave $\Leftrightarrow -f$ is convex [cite: 49]

21.1 Checking convexity for functions of one variable

A function $f : \mathbb{R} \to \mathbb{R}$ is convex if its second derivative is ≥ 0 everywhere. [cite: 49]

Example: $f(z) = z^2$ [cite: 50]

22 Checking convexity

22.1 Function of one variable

$F : \mathbb{R} \to \mathbb{R}$

- Value: number
- Derivative: number
- Second derivative: number

Convex if second derivative is always ≥ 0

22.2 Function of d variables

$F : \mathbb{R}^d \to \mathbb{R}$

- Value: number
- Derivative: d-dimensional vector
- Second derivative: $d \times d$ matrix

Convex if second derivative matrix is always positive semidefinite

22.3 First and second derivatives of multivariate functions

For a function $f : \mathbb{R}^d \to \mathbb{R}$

- the first derivative is a vector with d entries:

  $\nabla f(z) = \begin{pmatrix} \frac{\partial f}{\partial z_1} \\ \vdots \\ \frac{\partial f}{\partial z_d} \end{pmatrix}$

- the second derivative is a $d \times d$ matrix, the Hessian $H(z)$:

  $H_{jk} = \frac{\partial^2 f}{\partial z_j \partial z_k}$

23 Example

Find the second derivative matrix of $f(z) = ||z||^2$. [cite: 52]

23.1 When is a square matrix "positive"?

- A superficial notion: when all its entries are positive
- A deeper notion: when the quadratic function defined by it is always positive

Example:
$M = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ [cite: 53]

24 Positive semidefinite matrices

Recall: every square matrix M encodes a quadratic function:
$x \mapsto x^T M x = \sum_{i,j=1}^{d} M_{ij} x_i x_j$
(M is a $d \times d$ matrix and x is a vector in $\mathbb{R}^d$)
A symmetric matrix M is positive semidefinite (psd) if:
$x^T M x \geq 0$ for all vectors x
A symmetric matrix M is positive semidefinite (psd) if:
$x^T M x \geq 0$ for all vectors x
We saw that $\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ is PSD. [cite: 54]
What about $\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$? [cite: 55]

25

A symmetric matrix M is positive semidefinite (psd) if:
$x^T M x \geq 0$ for all vectors z
When is a diagonal matrix PSD? [cite: 56]
A symmetric matrix M is positive semidefinite (psd) if:
$x^T M x \geq 0$ for all vectors z
If M is PSD, must cM be PSD for a constant c? [cite: 57]

26

A symmetric matrix M is positive semidefinite (psd) if:
$x^T M x \geq 0$ for all vectors z
If M, N are of the same size and PSD, must $M + N$ be PSD? [cite: 58]

26.1 Checking if a matrix is PSD

A matrix M is PSD if and only if it can be written as $M = UU^T$ for some matrix U. [cite: 59]

Quick check: say $U \in \mathbb{R}^{r \times d}$ and $M = UU^T$

- M is square. [cite: 59]
- M is symmetric. [cite: 60]

Pick any $x \in \mathbb{R}^r$ Then
$x^T M x = x^T U U^T x = (x^T U)(U^T x)$
$= (U^T x)^T (U^T x)$
$= ||U^T x||^2 \geq 0$. [cite: 60]
Another useful fact: any covariance matrix is PSD. [cite: 60]

27 A hierarchy of square matrices

Square
$M \in \mathbb{R}^{d \times d}$
Symmetric
$M = M^T$
Positive semidefinite
$x^T M x \geq 0$ for all $x \in \mathbb{R}^d$ [cite: 61]

27.1 Second-derivative test for convexity

A function of several variables, $F(z)$, is convex if its second-derivative matrix $H(z)$ is positive semidefinite for all z. [cite: 61]

More formally:

Suppose that for $f : \mathbb{R}^d \to \mathbb{R}$, the second partial derivatives exist everywhere and are continuous functions of z. [cite: 62, 63] Then:

- $H(z)$ is a symmetric matrix
- f is convex $\Leftrightarrow$ $H(z)$ is positive semidefinite for all $z \in \mathbb{R}^d$ [cite: 63]

28 Example

Is $f(x) = ||x||^2$ convex? [cite: 64]

29 Example

Fix any vector $u \in \mathbb{R}^d$ Is this function $f : \mathbb{R}^d \to \mathbb{R}$ convex?

$f(z) = (u \cdot z)^2$ [cite: 64]

30 Least-squares regression

Recall loss function: for data points $(x^{(i)}, y^{(i)}) \in \mathbb{R}^d \times \mathbb{R}$,
$L(w) = \sum_{i=1}^{n}(y^{(i)} - (w \cdot x^{(i)}))^2$ [cite: 65]