Advanced Machine Learning Subsidary Notes

Lecture 12: Convexity

Adam Prügel-Bennett

April 28, 2020

1 Keywords

- Convex sets, convex functions, Jensen's inequality

2 Main Points

2.1 Convex Sets

- We are familiar geometrically with convex regions
- To define convexity we need to define an intermediate point $\boldsymbol{z} = a\,\boldsymbol{x} + (1-a)\,\boldsymbol{y}$
  - we requires $a \in [0, 1]$ (i.e. x is in the interval between 0 and 1) for $\boldsymbol{z}$ to be between $\boldsymbol{x}$ and $\boldsymbol{y}$
  - to define convexity we only need to have addition and scalar multiplication
- We can the define convexity in a very general way: a set $\mathcal{S}$ is convex if for every pair of points $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}$ and for every possible $a \in [0, 1]$ then $\boldsymbol{z} = a\,\boldsymbol{x} + (1-a)\,\boldsymbol{y} \in \mathcal{S}$
- We can apply convexity to sets of complicated objects
- For example the set of positive semi-definite matrices forms a convex set
  - This follows from the fact the sum of two positive semi-definite matrices is also positive semi-definite and if we multiply a positive semi-definite matrix by a positive number then the matrix is still positive semi-definite

2.2 Convex Functions

- We can define a function, $f(\boldsymbol{x})$, to be a *convex function* if for all pairs of points in the domain of the function and all $a \in [0, 1]$

$$f(a\,\boldsymbol{x} + (1-a)\,\boldsymbol{y}) \le a\,f(\boldsymbol{x}) + (1-a)\,f(\boldsymbol{y})$$

- This means that the function sits on or below the linear chord connecting any two points in the domain of the function
- The epigraph of a function is the area that lies on or above the functions
  - The epigraph of a convex function forms a convex region
  - If the epigraph of a function forms a convex region then the function is convex
- We can define *convex-down* or *concave* functions by inverting the constraint

$$f(a\,\boldsymbol{x} + (1-a)\,\boldsymbol{y}) \ge a\,f(\boldsymbol{x}) + (1-a)\,f(\boldsymbol{y})$$

  - for clarity I will sometimes refer to "convex" functions as *convex-up* functions

- convex-down functions have similar (but opposite) properties to convex up functions

- A function where for every pair of points and for any a such that $0 < a < 1$ (i.e. a lies strictly between 0 and 1) then if

$$f(a\,\boldsymbol{x} + (1-a)\,\boldsymbol{y}) < a\,f(\boldsymbol{x}) + (1-a)\,f(\boldsymbol{y})$$

then function is said to be *strictly convex*

- Linear functions $f(\boldsymbol{x}) = a\,\boldsymbol{x} + c$ are both convex-up and convex-down functions
  - For a function to be a strictly convex function it cannot have a linear section

- Convex functions lie on or above their tangent plane
  - The tangent plane to a function $f(\boldsymbol{x})$ at a point $\boldsymbol{x}_0$ is the plane orthogonal to the gradient, $\boldsymbol{\nabla} f(\boldsymbol{x}_0)$ that goes through the point $\boldsymbol{x}_0$

- A necessary and sufficient condition for a function to be convex is that its second derivative is non-negative or for multi-dimensional functions the Hessian is positive semi-definite
  - If the second derivative is positive (i.e. always greater than 0) or the Hessian is positive definite then the function is strictly positive

- Examples
  - Convex-up Functions
    * $f(x) = x^2$ is strictly convex since $f''(x) = 2 > 0$
    * $f(x) = x^{-2}$ is strictly convex since $f''(x) = 2\,x^{-4} > 0$
    * $f(x) = x^4$ is convex since $f''(x) = 12\,x^2 \geq 0$
    * $f(x) = \mathrm{e}^{c\,x}$ is strictly convex for all c as $f''(x) = c^2\,\mathrm{e}^{c\,x} > 0$
    * $f(\boldsymbol{x}) = \|\boldsymbol{x}\|^2$ is strictly convex since $\mathbf{H}(x) = \mathbf{I} \succ 0$
    * $f(x) = |x|$ is convex since for $a \in [0,1]$ we have $|a\,x + (1-a)\,y| \leq a\,|x| + (1-a)\,|y|$ with equality only when $x\,y \geq 0$
  - Convex-down Functions
    1. $f(x) = -x^2$ is strictly convex-down since $f''(x) = -2 < 0$
    2. $f(x) = \sqrt{x}$ (for $x > 0$) is strictly convex-down since $f''(x) = -x^{-3/2}/4 < 0$
    3. $f(x) = \log(x)$ is strictly convex-down since $f''(x) = -1/x^2 < 0$

- A function $f(\boldsymbol{x})$ that is constrained to a convex domain ($\boldsymbol{x} \in \mathcal{S}$, where $\mathcal{S}$ is a convex set) is convex in that domain if for all pairs $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}$ and all $a \in [0,1]$ we have

$$f(a\,\boldsymbol{x} + (1-a)\,\boldsymbol{y}) \leq a\,f(\boldsymbol{x}) + (1-a)\,f(\boldsymbol{y})$$

  - This is just a more precise definition of a convex function
  - Note that by limiting the domain of a function some non-convex functions may be convex over that domain
    * e.g. $\cos(x)$ is convex in the interval $[-\pi/2, 3\pi/3]$
  - A convex function constrained to lie in a convex set will still be convex
  - Any combination of linear constraints will form a convex set
  - Therefore convex functions restricted to satisfy linear constraints will be convex

- The minimum of a convex function will form a convex set
  - There can be no local minima
  - For a strictly convex function the minimum will be unique

- The sum of convex functions will be convex
- **Linear regression**
  - The loss function of linear regression is convex

$$L(\boldsymbol{w}) = \|\mathbf{X}\boldsymbol{w} - \boldsymbol{y}\|^2$$

    * The Hessian is $\mathbf{X}^\mathsf{T}\mathbf{X}$ which is positive semi-definite which is a sufficient condition for $L(\boldsymbol{w})$ to be convex
  - Both the L_2 regulariser and the L_1 regulariser are convex
  - The L_2 regulariser is strictly convex so there will be a unique solution
  - Many machine learning algorithms are chosen because they involve minimising a convex function leading to a unique minimum

2.3 Jensen's Inequality

- For any convex-up function, if $\boldsymbol{x}$ is a random variable then

$$\mathbb{E}[f(\boldsymbol{x})] \geq f(\mathbb{E}[\boldsymbol{x}])$$

  - $\mathbb{E}[\cdots]$ denotes the expectation
- For any convex-down function

$$\mathbb{E}[f(\boldsymbol{x})] \leq f(\mathbb{E}[\boldsymbol{x}])$$

- These are known as *Jensen's Inequality*
- **Proof**
  - We can prove this starting from the fact that $f(\boldsymbol{x})$ lies above the tangent plane at any point

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*) + (\boldsymbol{x} - \boldsymbol{x}^*)^\mathsf{T} \boldsymbol{\nabla} f(\boldsymbol{x}^*)$$

  - This has to be true at the point $\boldsymbol{x}^* = \mathbb{E}[\boldsymbol{x}]$

$$f(\boldsymbol{x}) \geq f(\mathbb{E}[\boldsymbol{x}]) + (\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])^\mathsf{T} \boldsymbol{\nabla} f(\mathbb{E}[\boldsymbol{x}])$$

  - Taking expectations of both sides of the equation

$$\mathbb{E}[f(\boldsymbol{x})] \geq f(\mathbb{E}[\boldsymbol{x}]) + (\mathbb{E}[\boldsymbol{x}] - \mathbb{E}[\boldsymbol{x}])^\mathsf{T} \boldsymbol{\nabla} f(\mathbb{E}[\boldsymbol{x}]) = f(\mathbb{E}[\boldsymbol{x}]) \qquad \square$$

- Using Jensen's Inequality
  - Consider the strictly convex function $f(\boldsymbol{x}) = x^2$ by Jensen's inequality

$$\mathbb{E}\left[x^2\right] \geq \mathbb{E}[x]^2$$

    * or $\mathbb{E}\left[x^2\right] - \mathbb{E}[x]^2 \geq 0$
      · the left-hand side is the variance so we see variances are non-negative
      · because $f(\boldsymbol{x}) = x^2$ is strictly convex we only get equality where $\boldsymbol{x}$ doesn't vary at all
  - Consider the Kullback-Liebler (KL) divergence defined for discrete probability probability distributions defined over the same range as

$$\mathcal{KL}(f\|g) = -\sum_i f_i \log\left(\frac{g_i}{f_i}\right)$$

* This is often used to measure how different distribution are from each other
* Note if $g_i = f_i$ then $\mathcal{KL}(f\|g) = 0$ since $\log(1) = 0$
* Now we can use Jensen's inequality to show that $\mathcal{KL}(f\|g) \geq 0$

$$
\begin{aligned}
\mathrm{KL}(f\|g) &= -\sum_i f_i \log\left(\frac{g_i}{f_i}\right) = -\mathbb{E}_f\left[\log\left(\frac{g_i}{f_i}\right)\right] \\
&\geq -\log\left(\mathbb{E}_f\left[\frac{g_i}{f_i}\right]\right) \\
&= -\log\left(\sum_i f_i \frac{g_i}{f_i}\right) = -\log\left(\sum_i g_i\right) = -\log(1) = 0
\end{aligned}
$$

  · Here we are assuming we have random variable that take values $X_i = g_i/f_i$ that occur with probability f_i
  · The KL-divergence is therefore equal to $\mathbb{E}[-\log(X_i)]$
  · Since $-\log(x)$ is convex up we have by Jensen's inequality that the KL-divergences is greater than or equal to $-\log(\mathbb{E}[X_i]) = -\log(\sum_i f_i X_i)$
  · But $X_i = g_i/f_i$ so the KL-divergence is greater than $-\log(\sum_i g_i)$
  · But g_i is a probability so $\sum_i g_i = 1$ giving us our result
* This is known as the Gibbs' inequality after the mathematical physicist, J. Willard Gibbs, (founder of modern statistical mechanics) who first proved it
* We often use KL-divergences when we want to choose the parameters of one probability distribution so that it approximates a second probability distribution

3 Exercises

3.1 Positive quadrant

- Prove that the set of vectors with non-negative elements form a convex set

3.2 Inverse of Convex Functions

1. Use the chain rule to compute the second derivative of $f(g(x))$
2. If $g(x) = f^{-1}(x)$ show that the second derivative of $f(g(x))$ vanishes
3. Use these results to derive an identity for the second derivative of $f^{-1}(x)$
4. Derive a condition for $f^{-1}(x)$ to be a convex-down function given that $f(x)$ is convex-up
5. Use this to show
   (a) $\sqrt{x}$ is a convex-down function
   (b) $\log(x)$ is a convex-down function

3.3 Cumulant Generating Function

- Here is something a bit harder (which you don't need to know)
- The cumulant generating function of a probability distribution $p(x)$ is defined as

$$
G(\lambda) = \log\left(\mathbb{E}\left[e^{\lambda x}\right]\right)
$$

  – the expectation is over the random variable x drawn from $p(x)$

- It is called the cumulant generating function because it we take then n^{th} derivative and set λ to zero we obtain the n^{th} cumulant (i.e. $\kappa_n = G^{(n)}(0)$)
- The first cumulant is the mean, the second the variance while the third and forth are proportional to the skewness and kurtosis
- Cumulant generating functions appear a lot when you work with probabilities, but go beyond this course
- Nevertheless let's show they are convex
  1. Find the second derivative
  2. Show that if $p(x)$ is a probability distribution then $q(x) = p(x)\,\mathrm{e}^{\lambda x}/\mathbb{E}\left[\mathrm{e}^{\lambda x}\right]$ is also a probability distribution
  3. Hence show that the cumulant generating function is convex
- See answers

4 Answers

4.1 Positive quadrant

- Let $\mathcal{P}$ be the set of vectors with non-negative elements
- If $\boldsymbol{x} \in \mathcal{P}$ then if $c \geq 0$ we have $\boldsymbol{v} = c\,\boldsymbol{x} \in \mathcal{P}$ since each element of $\boldsymbol{v}$ will be non-negative (i.e. $v_i = c\,x_i \geq 0$)
- Also for any two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{P}$ clearly $\boldsymbol{w} = \boldsymbol{x} + \boldsymbol{y} \in \mathcal{P}$ since $w_i = x_+ y_i$
- Thus for any two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{P}$ and any $a \in \{0, 1\}$ the vector $\boldsymbol{z} = a\,\boldsymbol{x} + (1-a)\,\boldsymbol{y}$ will be in $\mathcal{P}$

4.2 Inverse of Convex Functions

1. Taking derivatives
$$\frac{\mathrm{d}^2 f(g(x))}{\mathrm{d}x^2} = \frac{\mathrm{d} f'(g(x))\,g'(x)}{\mathrm{d}x} = f''(g(x))\,(g'(x))^2 + f'(g(x))\,g''(x)$$
2. If $g(x) = f^{-1}(x)$ then $f(g(x)) = x$ and the second derivative vanishes
3. Using 1. and 2. we find (writing $f^{-1}(x)$ as $g(x)$)
$$g''(x) = -\frac{f''(g(x))\,(g'(x))^2}{f'(g(x))}$$
4. If $f(x)$ is convex then $f''(y) \geq 0$ for any y (including $y = f^{-1}(x)$) also $(g'(x))^2 \geq 0$ so for the inverse of $f(x)$ to be convex down we require $f'(f^{-1}(x)) > 0$
5. Use this to show
   (a) Let $f(x) = x^2$, so that $f''(x) = 2 > 0$ and $f'(y) = y$ which is non-negative if $y \geq 0$, but $f^{-1}(x) = \sqrt{x} > 0$ so $f'(f^{-1}(x)) \geq 0$ and consequently $\sqrt{x}$ is convex-down
   (b) Let $f(x) = \exp(x)$, so that $f''(x) = \exp(x) > 0$. But $f'(y) = \exp(y) > 0$ for all y so $f'(f^{-1}(x)) > 0$ which is sufficient to show $f^{-1}(x) = \log(x)$ is a convex-down function

4.3 Cumulant Generating Function

1. If $G(\lambda) = \log\!\left(\mathbb{E}\left[e^{\lambda x}\right]\right)$ then

$$G'(\lambda) = \frac{\mathbb{E}\left[x\,e^{\lambda x}\right]}{\mathbb{E}\left[e^{\lambda x}\right]}$$

   and

$$G''(\lambda) = \frac{\mathbb{E}\left[x^2\,e^{\lambda x}\right]}{\mathbb{E}\left[e^{\lambda x}\right]} - \frac{\mathbb{E}\left[x\,e^{\lambda x}\right]^2}{\mathbb{E}\left[e^{\lambda x}\right]^2}$$

2.
   - Now if $p(x)$ is a probability distribution is will be non-negative for all x and sum or integrate to 1
   - But then $q(x) = p(x)\,e^{\lambda x}/\mathbb{E}\left[e^{\lambda x}\right]$ will be non-negative as $e^{\lambda x} > 0$ and $\mathbb{E}\left[e^{\lambda x}\right] > 0$ (the expectation of positive quantities will be positive)
   - But

$$\int q(x)\,\mathrm{d}x = \frac{1}{\mathbb{E}\left[e^{\lambda x}\right]}\int p(x)\,e^{\lambda x}\,\mathrm{d}x = \frac{\mathbb{E}\left[e^{\lambda x}\right]}{\mathbb{E}\left[e^{\lambda x}\right]} = 1$$

   - So $q(x)$ is non-negative and normalised so is a well defined probability distribution

3. Using the result of 1. and 2

$$G''(\lambda) = \frac{\mathbb{E}_p\left[x^2\,e^{\lambda x}\right]}{\mathbb{E}_p\left[e^{\lambda x}\right]} - \frac{\mathbb{E}_p\left[x\,e^{\lambda x}\right]^2}{\mathbb{E}_p\left[e^{\lambda x}\right]^2} = \mathbb{E}_g\left[x^2\right] - \mathbb{E}_g\left[x\right]^2 \geq 0$$

   - since variances are non-negative