Uma Introdução a Geometria da Informação

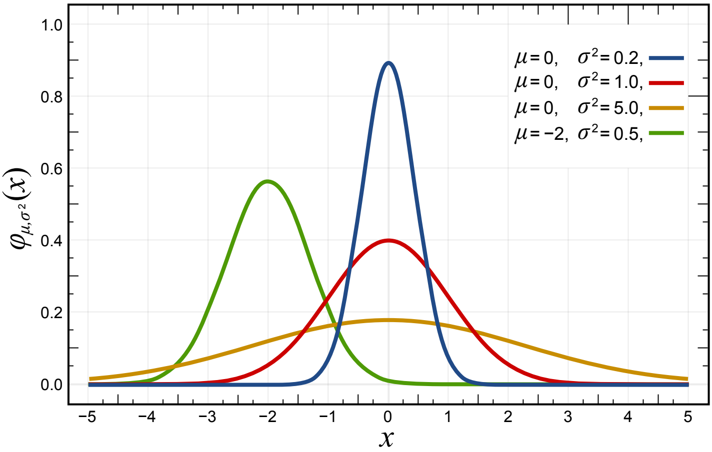
Fábio C. C. Meneghetti

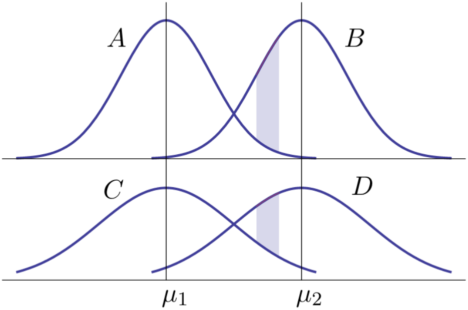
Instituto de Matemática, Estatística e Computação Científica
Universidade Estadual de Campinas

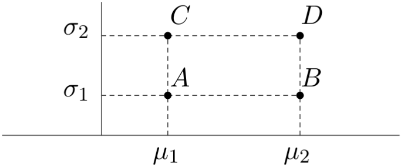
28 de maio de 2021

1. Motivação
2. Geometria Riemanniana
3. Variedades estatísticas
4. Modelos estatísticos

Distribuição normal (gaussiana): $f_{\mu,\sigma} = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\|x-\mu\|^2}{2\sigma^2}\right)$.




A
B
C
D
μ_1
μ_2


σ_2
σ_1
C
D
A
B
μ_1
μ_2

Definir uma estrutura de variedade Riemanniana em $M = \left\{p_\xi : \xi \in \Theta\right\}$ nos permite:

- Entender distâncias entre distribuições como geodésicas;
- Entender o quão sensível p_ξ é aos parâmetros ξ;
- Entender como fazer otimização sobre M, o que é essencial para decidir qual distribuição é mais apropriada para um dado problema (inferência estatística).

Variedade Riemanniana

Uma **variedade Riemanniana** é uma variedade diferenciável M munida de um produto interno $g_p = \langle \cdot, \cdot \rangle_p$ suave em T_pM.

Variedade Riemanniana

Uma **variedade Riemanniana** é uma variedade diferenciável M munida de um produto interno $g_p = \langle \cdot, \cdot \rangle_p$ suave em T_pM.

Denote por $\mathfrak{X}(M)$ o conjunto de campos de vetores suaves $X\colon M \to TM$.

Definição 2.1

Uma conexão é uma função $\nabla\colon \mathfrak{X}(M) \times \mathfrak{X}(M) \to \mathfrak{X}(M)$ que satisfaz:

1. $\nabla_{fX} Y = f \nabla_X Y$ ($C^\infty(M)$-linear na 1ª entrada)
2. $\nabla_X fY = X(f)Y + f\nabla_X Y$ (regra de Leibniz na 2ª entrada)

Variedade Riemanniana

Uma **variedade Riemanniana** é uma variedade diferenciável M munida de um produto interno $g_p = \langle\cdot,\cdot\rangle_p$ suave em T_pM.

Denote por $\mathfrak{X}(M)$ o conjunto de campos de vetores suaves $X\colon M \to TM$.

Definição 2.1

Uma conexão é uma função $\nabla\colon \mathfrak{X}(M) \times \mathfrak{X}(M) \to \mathfrak{X}(M)$ que satisfaz:

1. $\nabla_{fX}Y = f\nabla_X Y$ ($C^\infty(M)$-linear na 1ª entrada)
2. $\nabla_X fY = X(f)Y + f\nabla_X Y$ (regra de Leibniz na 2ª entrada)

Símbolos de Christoffel (locais): $\nabla_{\partial_i}\partial_j = \sum_k \Gamma_{ij}^k \partial_k$

Uma conexão é importante por permitir definir conceitos como transporte paralelo ao longo de uma curva γ, e geodésicas.

- Um campo X ao longo de γ é parelelo se $\nabla_{\dot{\gamma}} X \equiv 0$.
- Uma geodésica é uma curva autoparalela ($\nabla_{\dot{\gamma}} \dot{\gamma} \equiv 0$).

Uma conexão é importante por permitir definir conceitos como transporte paralelo ao longo de uma curva γ, e geodésicas.

- Um campo X ao longo de γ é parelelo se $\nabla_{\dot{\gamma}} X \equiv 0$.
- Uma geodésica é uma curva autoparalela ($\nabla_{\dot{\gamma}} \dot{\gamma} \equiv 0$).

Teorema 2.2 (Conexão de Levi-Civita)

Seja (M, g) variedade Riemanniana. Então existe única conexão ${}^{\mathsf{LC}}\nabla$ que é:

1. Compatível com a métrica: $X\langle Y, Z\rangle = \langle {}^{\mathsf{LC}}\nabla_X Y, Z\rangle + \langle Y, {}^{\mathsf{LC}}\nabla_X Z\rangle$.
2. Livre de torção: $T(X, Y) = {}^{\mathsf{LC}}\nabla_X Y - {}^{\mathsf{LC}}\nabla_Y X - [X, Y] = 0$.

1. Motivação

2. Geometria Riemanniana

3. Variedades estatísticas

4. Modelos estatísticos

Conexões duais

Consideremos aqui conexões sempre livres de torção, mas não necessariamente de Levi-Civita.

Conexões duais

Consideremos aqui conexões sempre livres de torção, mas não necessariamente de Levi-Civita.

Definição 3.1

Dada uma conexão ∇, dizemos que uma outra conexão ∇^* é **dual** (ou conjugada) com respeito à métrica g se

$$X\langle Y, Z\rangle = \langle \nabla_X Y, Z\rangle + \langle Y, \nabla^*_X Z\rangle$$

Propriedades:

1. O dual é **único**.

Propriedades:

1. O dual é **único**.
2. $({}^{\mathsf{LC}}\nabla)^* = {}^{\mathsf{LC}}\nabla$ (Conexão de Levi-Civita é autodual).

Propriedades:

1. O dual é **único**.
2. $({}^{\mathsf{LC}}\nabla)^* = {}^{\mathsf{LC}}\nabla$ (Conexão de Levi-Civita é autodual).
3. $\frac{1}{2}(\nabla + \nabla^*) = {}^{\mathsf{LC}}\nabla$ para qualquer par de conexões duais.

Propriedades:

1. O dual é **único**.
2. $\left({}^{\mathsf{LC}}\nabla\right)^* = {}^{\mathsf{LC}}\nabla$ (Conexão de Levi-Civita é autodual).
3. $\frac{1}{2}\left(\nabla + \nabla^*\right) = {}^{\mathsf{LC}}\nabla$ para qualquer par de conexões duais.
4. Transporte paralelo dual preserva a métrica:

$$\langle v, w\rangle_{\gamma(0)} = \left\langle \prod_{\gamma}^{\nabla} v, \prod_{\gamma}^{\nabla^*} w \right\rangle_{\gamma(1)}$$

Teorema 3.2 (Fundamental de Geometria da Informação)

(M, g, ∇) tem curvatura contante $\kappa \iff (M, g, \nabla^*)$ tem curvatura contante κ.

- Se as $\kappa = 0$, dizemos que a variedade é **dualmente plana** (isso vale em particular quando existe um sistema de coordenadas onde $\Gamma_{i,j}^k$ se anulam).

Teorema 3.2 (Fundamental de Geometria da Informação)

(M, g, ∇) tem curvatura contante $\kappa \iff (M, g, \nabla^*)$ tem curvatura contante κ.

- Se as $\kappa = 0$, dizemos que a variedade é **dualmente plana** (isso vale em particular quando existe um sistema de coordenadas onde $\Gamma_{i,j}^k$ se anulam).

Definição 3.3

A tripla (g, ∇, ∇^*) é chamada de **estrutura dualística** para a variedade M.

Construção equivalente

- Frequentemente (M, g, ∇, ∇^*) é chamada de *variedade estatística*.

[1] Steffen L. Lauritzen. "Chapter 4: Statistical Manifolds". Em: *Institute of Mathematical Statistics Lecture Notes - Monograph Series Differential geometry in statistical inference* (1987), pp. 163–216. DOI: 10.1214/lnms/1215467061.

Construção equivalente

- Frequentemente (M, g, ∇, ∇^*) é chamada de *variedade estatística*.
- Mas Lauritzen[1] define **variedade estatística** como (M, g, C), onde C é um 3-tensor covariante simétrico, isto é,

$$C(X, Y, Z) = C(Y, X, Z) = C(Y, Z, X)$$

[1] Steffen L. Lauritzen. "Chapter 4: Statistical Manifolds". Em: *Institute of Mathematical Statistics Lecture Notes - Monograph Series Differential geometry in statistical inference* (1987), pp. 163–216. DOI: 10.1214/lnms/1215467061.

Construção equivalente

- Frequentemente (M, g, ∇, ∇^*) é chamada de *variedade estatística*.
- Mas Lauritzen[1] define **variedade estatística** como (M, g, C), onde C é um 3-tensor covariante simétrico, isto é,

$$C(X,Y,Z) = C(Y,X,Z) = C(Y,Z,X)$$

- As definições são consideradas equivalentes:
  - $C(X,Y,Z) = \langle \nabla_X Y - \nabla^*_X Y, Z \rangle$;

[1] Steffen L. Lauritzen. "Chapter 4: Statistical Manifolds". Em: *Institute of Mathematical Statistics Lecture Notes - Monograph Series Differential geometry in statistical inference* (1987), pp. 163–216. DOI: 10.1214/lnms/1215467061.

Construção equivalente

- Frequentemente (M, g, ∇, ∇^*) é chamada de *variedade estatística*.
- Mas Lauritzen[1] define **variedade estatística** como (M, g, C), onde C é um 3-tensor covariante simétrico, isto é,

$$C(X, Y, Z) = C(Y, X, Z) = C(Y, Z, X)$$

- As definições são consideradas equivalentes:
  - $C(X, Y, Z) = \langle \nabla_X Y - \nabla^*_X Y, Z \rangle$;
  - A partir de C construímos uma estrutura dualística $\left(\nabla^\alpha, \nabla^{-\alpha}\right)$, $\alpha \in \mathbb{R}$:

$$\langle \nabla^\alpha_X Y, Z \rangle := \langle {}^{\mathsf{LC}}\nabla_X Y, Z \rangle - \tfrac{\alpha}{2} C(X, Y, Z).$$

[1] Steffen L. Lauritzen. “Chapter 4: Statistical Manifolds”. Em: *Institute of Mathematical Statistics Lecture Notes - Monograph Series Differential geometry in statistical inference* (1987), pp. 163–216. DOI: 10.1214/lnms/1215467061.

Variedades estatísticas a partir de divergências

Notação: $\partial_{i,\cdot} f(x,y) = \frac{\partial}{\partial x_i} f(x,y)$.

Variedades estatísticas a partir de divergências

Notação: $\partial_{i,\cdot} f(x, y) = \frac{\partial}{\partial x_i} f(x, y)$.

Definição 3.4

Uma **divergência** $D\colon M \times M \to [0, \infty)$ com respeito a uma carta φ é uma função C^3 satisfazendo: $(p, q \in \mathrm{dom}\varphi)$

1. $D(p : q) \geq 0$, com igualdade $\iff$ $p = q$,
2. $\partial_{i,\cdot} D(p : q)\big|_{p=q} = \partial_{\cdot,j} D(p : q)\big|_{p=q} = 0$ para todo i, j,
3. $-\,\partial_{\cdot,i}\partial_{\cdot,j} D(p : q)\big|_{p=q}$ é positiva-definida.

Variedades estatísticas a partir de divergências

Notação: $\partial_{i,\cdot} f(x,y) = \frac{\partial}{\partial x_i} f(x,y)$.

Definição 3.4

Uma **divergência** $D\colon M \times M \to [0,\infty)$ com respeito a uma carta φ é uma função C^3 satisfazendo: $(p, q \in \mathrm{dom}\varphi)$

1. $D(p:q) \geq 0$, com igualdade $\iff p = q$,
2. $\partial_{i,\cdot} D(p:q)\big|_{p=q} = \partial_{\cdot,j} D(p:q)\big|_{p=q} = 0$ para todo i, j,
3. $-\partial_{\cdot,i}\partial_{\cdot,j} D(p:q)\big|_{p=q}$ é positiva-definida.

- Divergência dual: $D^*(p:q) := D(q:p)$.

Toda divergência dá origem a uma estrutura dual:

- ${}^{D}g_{ij} := \partial_{i,j} D(p:q)\big|_{p=q} = {}^{D^*}g_{ij}$,

Toda divergência dá origem a uma estrutura dual:

- ${}^{D}g_{ij} := \partial_{i,j} D(p:q)\big|_{p=q} = {}^{D^*}g_{ij}$,
- ${}^{D}\Gamma_{ij}^{k} = \partial_{ij,k} D(p:q)\Big|_{p=q}$ (coordenadas locais da conexão).

Toda divergência dá origem a uma estrutura dual:

- ${}^{D}g_{ij} := \partial_{i,j} D(p:q)\big|_{p=q} = {}^{D^*}g_{ij}$,
- ${}^{D}\Gamma_{ij}^{k} = \partial_{ij,k} D(p:q)\Big|_{p=q}$ (coordenadas locais da conexão).

Propositção 3.5

Se ${}^{D}\nabla$ é a conexão definida por ${}^{D}\Gamma_{ij}^{k}$, temos:

$$ {}^{D^*}\nabla = \left({}^{D}\nabla\right)^* $$

- Toda variedade estatística vem de uma divergência.[2]

[2] Takao Matumoto. "Any statistical manifold has a contrast function — on the C^3-functions taking the minimum at the diagonal of the product manifold". Em: *Hiroshima Mathematical Journal* 23.2 (1993), pp. 327–332. DOI: 10.32917/hmj/1206128255.

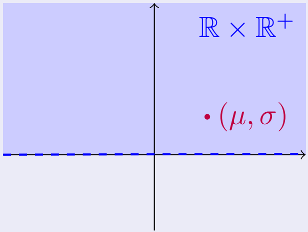
Modelos estatísticos

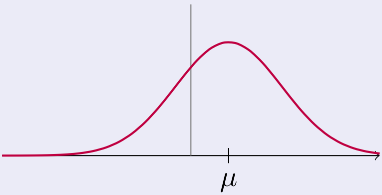
Um modelo estatístico é uma família $\mathcal{M} = \left\{ p_\xi \right\}_{\xi \in \Theta}$, onde Θ é aberto de $\mathbb{R}^n$, e cada p_ξ é uma função densidade de probabilidade.

- Podemos tomar p_ξ como sendo funções $\mathbb{R}^n \to \mathbb{R}^+$ contínuas, integráveis com $\int_{\mathbb{R}^n} p_\xi \, \mathrm{d}x = 1$.

Exemplo 4.1

Tome $\mathcal{M} = \left\{ p_{\mu,\sigma} = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\|x-\mu\|^2}{2\sigma^2} \right) \,\middle|\, (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+ \right\}$ a família das distribuições normals univariadas.





Métrica da informação de Fisher

A métrica Riemanniana "padrão" em modelos estatísticos é a **métrica da informação de Fisher**:

$$\begin{aligned} g_{ij} &= \int_{\mathbb{R}^n} p_\xi(x) \frac{\partial \log(p_\xi(x))}{\partial \xi_i} \frac{\partial \log(p_\xi(x))}{\partial \xi_j} \, \mathrm{d}x \\ &=^{3} - \int_{\mathbb{R}^n} p_\xi(x) \frac{\partial^2 \log(p_\xi(x))}{\partial \xi_i \partial \xi_j} \, \mathrm{d}x \end{aligned}$$

[3] Sob algumas condições de regularidade.
[3] Sob algumas condições de regularidade.

Métrica da informação de Fisher

A métrica Riemanniana "padrão" em modelos estatísticos é a **métrica da informação de Fisher**:

$$g_{ij} = \int_{\mathbb{R}^n} p_\xi(x) \frac{\partial \log(p_\xi(x))}{\partial \xi_i} \frac{\partial \log(p_\xi(x))}{\partial \xi_j} \, \mathrm{d}x$$
$$= ^{3} - \int_{\mathbb{R}^n} p_\xi(x) \frac{\partial^2 \log(p_\xi(x))}{\partial \xi_i \partial \xi_j} \, \mathrm{d}x$$

Obs: Podemos escrever também $g_{ij} = \mathrm{E}_{p_\xi}[\partial_i \ell_\xi \partial_j \ell_\xi]$, onde $\ell_\xi(x) = \log(p_\xi(x))$ é a função log-verossimilhança.

[3] Sob algumas condições de regularidade.

[3] Sob algumas condições de regularidade.

Teorema de Chentsov

Teorema 4.2

A métrica de Fisher é a **única** métrica Riemanniana (a menos de constante) invariante por estatísticas suficientes.[a]

[a] Nihat Ay et al. "Information geometry and sufficient statistics". Em: *Probability Theory and Related Fields* 162.1-2 (jun. de 2014), pp. 327–364. ISSN: 1432-2064. DOI: 10.1007/s00440-014-0574-8.

Teorema de Chentsov

Teorema 4.2

A métrica de Fisher é a **única** métrica Riemanniana (a menos de constante) invariante por estatísticas suficientes.[a]

[a] Nihat Ay et al. "Information geometry and sufficient statistics". Em: *Probability Theory and Related Fields* 162.1-2 (jun. de 2014), pp. 327–364. ISSN: 1432-2064. DOI: 10.1007/s00440-014-0574-8.

- Informalmente, uma estatística suficiente é uma mudança do espaço de parâmetros ξ sem perda de informações sobre a variável aleatória correspondente a p_ξ.

Entropia relativa

A métrica de Fisher define uma variedade estatística através da **divergência de Kullback-Leibler** (ou *entropia relativa*), uma das principais divergências em teoria da informação:

$$D_{\mathsf{KL}}(p : q) = \int_{\mathbb{R}^n} p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x$$

Fazendo a construção mostrada anteriormente, obtemos com a entropia relativa, obtemos uma variedade estatística **dualmente plana**:

- $g_{ij} = {}^{D_{\mathsf{KL}}}g_{ij} = \partial_{i,j} D_{\mathsf{KL}}(p_\xi : p_{\xi'})\Big|_{\xi=\xi'}$
- ${}^{D_{\mathsf{KL}}}\Gamma_{ij}^{k} = 0$
- ${}^{D^*_{\mathsf{KL}}}\Gamma_{ij}^{k} = 0$

Generalidade

- Em 2005, Hông Vân Lê mostrou que toda variedade estatística abstrata pode ser descrita como uma família de distribuições de probabilidade com a métrica de Fisher[4].
- A demonstração usa resultados importantes de geometria, como o teorema da imersão de Nash.

[4] Hông Vân Lê. "Statistical manifolds are statistical models". Em: *Journal of Geometry* (2006). DOI: 10.1007/s00022-005-0030-0.

Métrica de Fisher-Rao

- A métrica de Fisher induz uma métrica sobre o modelo $\mathcal{M}$, dada pelas geodésicas minimizantes.

Métrica de Fisher-Rao

- A métrica de Fisher induz uma métrica sobre o modelo $\mathcal{M}$, dada pelas geodésicas minimizantes.
- Uma geodésica pode ser descrita pelas equações de Euler-Lagrange, como uma função $\gamma\colon [0,1] \to \mathcal{M}$ que satisfaz:

$$\frac{\mathrm{d}^2 x_k}{\mathrm{d}t^2} \sum_{i,j} \Gamma_{ij}^k \frac{\mathrm{d}x_i}{\mathrm{d}t} \frac{\mathrm{d}x_j}{\mathrm{d}t} = 0.$$

Métrica de Fisher-Rao

- A métrica de Fisher induz uma métrica sobre o modelo $\mathcal{M}$, dada pelas geodésicas minimizantes.
- Uma geodésica pode ser descrita pelas equações de Euler-Lagrange, como uma função $\gamma\colon [0,1] \to \mathcal{M}$ que satisfaz:

$$\frac{\mathrm{d}^2 x_k}{\mathrm{d}t^2} \sum_{i,j} \Gamma_{ij}^k \frac{\mathrm{d}x_i}{\mathrm{d}t} \frac{\mathrm{d}x_j}{\mathrm{d}t} = 0.$$

- A distância de Fisher-Rao é a métrica geodésica em $\mathcal{M}$, dada por

$$d_{\mathsf{F}}(p_\xi, p_\theta) = \inf_\gamma \int_0^1 \|\gamma'(t)\| \, \mathrm{d}t,$$

onde $\gamma(0) = p_\xi$, $\gamma(1) = p_\theta$.

Exemplo 4.3

Tomando $\mathcal{M}$ do exemplo anterior, podemos calcular[a] a matriz de Fisher como

$$[g_{ij}(\mu,\theta)] = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}$$

- A geometria definida por esta métrica coincide com a **geometria hiperbólica** sobre o meio-plano de Poincaré[b].

[a] Sueli I.R. Costa, Sandra A. Santos e João E. Strapasson. "Fisher information distance: A geometrical reading". Em: *Discrete Applied Mathematics* 197 (2015), pp. 59–69. ISSN: 0166-218X. DOI: https://doi.org/10.1016/j.dam.2014.10.004.

[b] Com uma pequena deformação por conta do termo 2 na última entrada.

Exemplo 4.3

- Nesta geometria, as geodésicas são dadas por retas verticais e por meias-elipses de excentricidade $1/\sqrt{2}$.

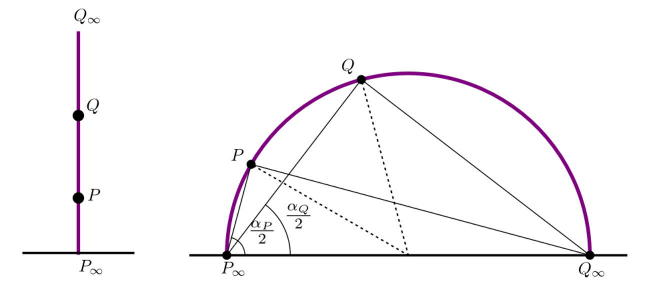


Figura: Geodésicas no meio-plano.

Mais detalhes no artigo [5].

Teoremas de Pitágoras duais

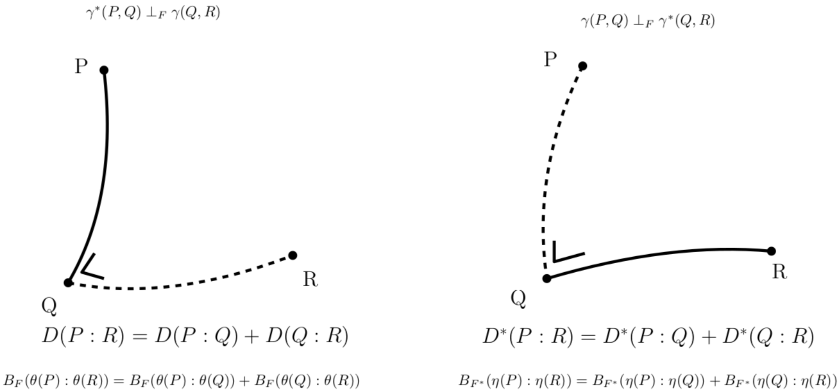


Figure 6. Dual Pythagorean theorems in a dually flat space.

Método do gradiente Riemanniano

De forma bem simplificada, podemos tratar o problema de aprendizado de máquina com *deep learning* da seguinte forma:

- Temos um conjunto de treinamento $\{(x_i, y_i)\}_{i=1}^{m}$, $x_i \in \mathbb{R}^n$, $y_i \in \{0, 1\}$;

Método do gradiente Riemanniano

De forma bem simplificada, podemos tratar o problema de aprendizado de máquina com *deep learning* da seguinte forma:

- Temos um conjunto de treinamento $\{(x_i, y_i)\}_{i=1}^{m}$, $x_i \in \mathbb{R}^n$, $y_i \in \{0, 1\}$;
- Temos uma família de funções parametrizadas $\left\{f_\xi\right\}_{\xi\in\Omega}$, com $f_\xi : \mathbb{R}^n \to \mathbb{R}$;

Método do gradiente Riemanniano

De forma bem simplificada, podemos tratar o problema de aprendizado de máquina com *deep learning* da seguinte forma:

- Temos um conjunto de treinamento $\{(x_i, y_i)\}_{i=1}^m$, $x_i \in \mathbb{R}^n$, $y_i \in \{0, 1\}$;
- Temos uma família de funções parametrizadas $\left\{f_\xi\right\}_{\xi\in\Omega}$, com $f_\xi \colon \mathbb{R}^n \to \mathbb{R}$;
- Queremos encontrar o $\bar{\xi}$ tal que $f_{\bar{\xi}}$ melhor se aproxime do conjunto de treinamento

- Para isso escolhe-se uma medida de dissimilaridade, chamada **função perda** $\mathcal{L}$. O problema do aprendizado consiste em encontrar ξ que minimize

$$\mathcal{L}_\xi = \mathcal{L}\left(f_\xi(\mathbf{x}), \mathbf{y}\right), \qquad \mathbf{x} = [x_i],\ \mathbf{y} = [y_i]$$

Por exemplo, podemos ter $\mathcal{L}(p, q) = D_{\mathsf{KL}}(p, q) = \sum_i p_i \log \frac{p_i}{q_i}$

- Para isso escolhe-se uma medida de dissimilaridade, chamada **função perda** $\mathcal{L}$. O problema do aprendizado consiste em encontrar ξ que minimize

$$\mathcal{L}_\xi = \mathcal{L}\left(f_\xi(\mathbf{x}), \mathbf{y}\right), \qquad \mathbf{x} = [x_i],\ \mathbf{y} = [y_i]$$

  Por exemplo, podemos ter $\mathcal{L}(p, q) = D_{\mathsf{KL}}(p, q) = \sum_i p_i \log \frac{p_i}{q_i}$

- Isso é quase sempre feito atreavés do método do gradiente:

$$\xi_{n+1} = \xi_n - \alpha \cdot \nabla_\xi \mathcal{L}_{\xi_n}$$

- O aprendizado de máquina, portanto, é um problema de otimização na variedade estatística do espaço de parâmetros.

- O aprendizado de máquina, portanto, é um problema de otimização na variedade estatística do espaço de parâmetros.
- Numa variedade Riemanniana, o gradiente Riemanniano $\nabla_M \mathcal{L}_\xi$ é definido como o vetor v que minimiza

$$\nabla_v \mathcal{L}(p) = \lim_{h \to 0} \frac{\mathcal{L}(\exp_p hv) - \mathcal{L}(p)}{h}.$$

- O aprendizado de máquina, portanto, é um problema de otimização na variedade estatística do espaço de parâmetros.
- Numa variedade Riemanniana, o gradiente Riemanniano $\nabla_M \mathcal{L}_\xi$ é definido como o vetor v que minimiza

$$\nabla_v \mathcal{L}(p) = \lim_{h \to 0} \frac{\mathcal{L}(\exp_p hv) - \mathcal{L}(p)}{h}.$$

- Se utilizamos a aproximação $\nabla_M \mathcal{L}_\xi \approx [g_{ij}(\xi)]^{-1} \nabla_\xi \mathcal{L}_\xi$ teremos o método do **gradiente Riemanniano**:

$$\xi_{n+1} = \xi_n - \alpha [g_{ij}(\xi)]^{-1} \nabla_\xi \mathcal{L}_\xi.$$

- No artigo[5], argumenta-se que, no espaço de parâmetros de redes neurais, o gradiente Riemanniano é o gradiente que realmente representa a direção de máxima descida.
- Ele é mostrado ser estatisticamente eficiente, e apresenta outras vantagens como a redução do "efeito platô"

[5] Shun-ichi Amari. "Natural Gradient Works Efficiently in Learning". Em: *Neural Computation* 10.2 (1998), pp. 251–276. DOI: 10.1162/089976698300017746.

Shun'ichi Amari. *Information Geometry and Its Applications*. Springer, 2016.

N. Ay, J. Jost, H.V. Lê e L. Schwachhöfer. *Information Geometry*. A Series of Modern Surveys in Mathematics. Springer International Publishing, 2017. ISBN: 9783319564784.

Sueli I.R. Costa, Sandra A. Santos e João E. Strapasson. "Fisher information distance: A geometrical reading". Em: *Discrete Applied Mathematics* 197 (2015), pp. 59–69. ISSN: 0166-218X. DOI: https://doi.org/10.1016/j.dam.2014.10.004.

Frank Nielsen. "An Elementary Introduction to Information Geometry". Em: *Entropy* 22.10 (2020), p. 1100. DOI: 10.3390/e22101100.

Dúvidas?