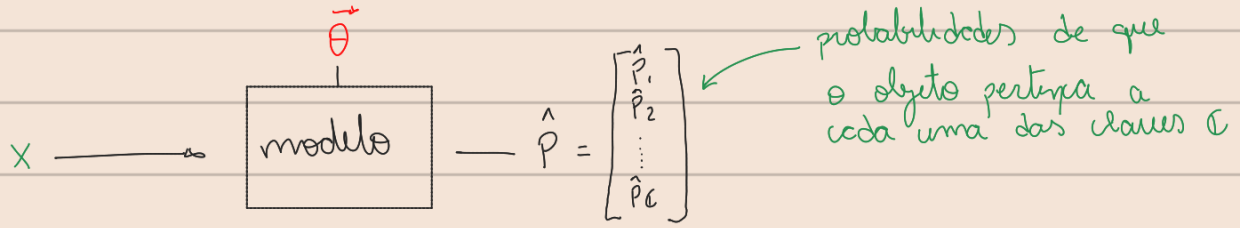
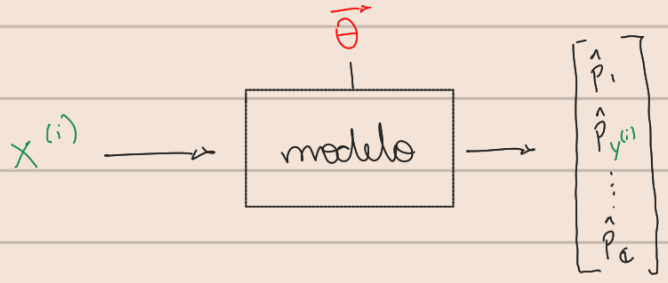
Treinando um classificador



Para um exemplo $(x^{(i)}, y^{(i)})$ qual é a chance, de acordo com o modelo, de que o objeto de features $x^{(i)}$ pertença a classe $y^{(i)}$?



$$\hat{P}_{y^{(i)}} = \text{Prob}\left(Y = y^{(i)} \mid X = x^{(i)}, \vec{\theta}\right)$$

- um modelo ruim discorda da realidade:
  - → $\text{Prob}\left(Y = y^{(i)} \mid X = x^{(i)}, \vec{\theta}_{RUIM}\right)$ vai ser baixa
- um modelo bom concorda com a realidade:
  - → $\text{Prob}\left(Y = y^{(i)} \mid X = x^{(i)}, \vec{\theta}_{BOM}\right)$ vai ser alta

IDEIA: Achar parâmetros $\vec{\theta}$ que maximizam a OPINIÃO DO MODELO acerca dos observados

$$\vec{\theta}_{ML} = \text{argmax Prob}\left[\text{targets do exemplo de treino} \mid \text{features do exemplo de treino}, \vec{\theta}\right] = \text{Prob}\left(\text{Dados} \mid \vec{\theta}\right)$$

verossimilhança (likelihood)

Construindo a formula de

$$\text{Prob}\left[\text{targets} \mid \text{features}, \theta\right]$$

· Premissas

1. os $(x^{(i)}, y^{(i)})$ vem da mesma distribuição $P(X,Y)$ (iid)
2. $(x^{(i)}, y^{(i)})$ são independentes

$$\text{Prob}\left[Y_1 = y^{(1)} \text{ e } Y_2 = y^{(2)} \text{ e } Y_m = y^{(m)} \mid X_1 = \vec{x}^{(1)} \text{ e } X_2 = \vec{x}^{(2)} \text{ e } X_m = \vec{x}^{(m)} \text{ e } \theta\right]$$

$$= \text{Prob}\left[Y = y^{(1)} \mid X = \vec{x}^{(1)}, \vec{\theta}\right] \cdot \text{Prob}\left[Y = y^{(2)} \mid X = \vec{x}^{(2)}, \vec{\theta}\right] \cdot \text{Prob}\left[Y = y^{(m)} \mid X = \vec{x}^{(m)}, \vec{\theta}\right]$$

$$= \text{Prob}\left[\text{targets} \mid \text{features}, \theta\right]$$

$$= \prod_{i=1}^{m} \hat{p}\, y^{(i)} \mid x^{(i)}, \vec{\theta}$$

$$\vec{\theta}_{ML} = \underset{\vec{\theta}}{\text{argmin}}\ \text{Prob}\left[\text{targets} \mid \text{features}, \theta\right]$$

$$= \underset{\vec{\theta}}{\text{argmax}}\ \log \text{Prob}\left[\text{targets} \mid \text{features}, \theta\right]$$

$$= \sum_{i=1}^{m} \log \hat{p}^{(i)}_{y^{(i)}}$$

$$= \sum_{i=1}^{m} \sum_{K=1}^{C} \left[y^{(i)} == K\right] \log\left(\hat{p}^{(i)}_{K}\left(x^{(i)}, \theta\right)\right)$$

$\left[y^{(i)} == K\right]$ → expressão booleana

Notação de Iverson

Entropia cruzada

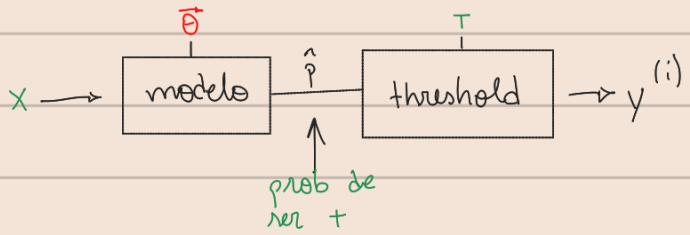
maximum likelihood:

$$\vec{\theta}^{\text{ótimo}} = \underset{\vec{\theta}}{\text{argmax}} \left\{ \sum_{i=1}^{m} \sum_{K=1}^{C} \left[y^{(i)} == K\right] \log\left(\hat{p}^{(i)}_{K}\left(x^{(i)}, \theta\right)\right) \right\}$$

$$\Updownarrow$$

$$\vec{\theta}^{\text{ótimo}} = \underset{\vec{\theta}}{\text{argmin}} \left\{ -\frac{1}{m} \sum_{i=1}^{m} \sum_{K=1}^{C} \left[y^{(i)} == K\right] \log\left(\hat{p}^{(i)}_{K}\left(x^{(i)}, \theta\right)\right) \right\} \quad * \text{ cross entropy}$$

Exemplo: classificador binário



$$\text{Loss} = -\frac{1}{m}\sum_{i=1}^{m}\sum_{K=1}^{C}\left[y^{(i)} == K\right]\log\left(\hat{p}_K^{(i)}\right)$$

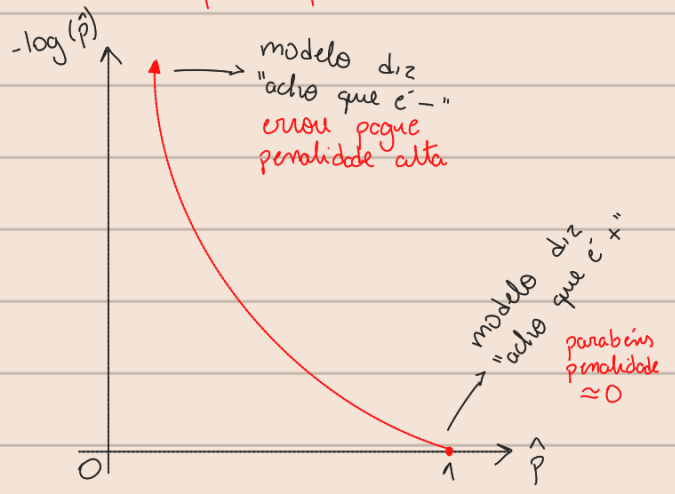
$$= -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)} == 0\right]\log\left(\hat{p}_K^{(i)}\right) + -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)} == 1\right]\log\left(\hat{p}_K^{(i)}\right)$$

$$= -\frac{1}{m}\left[(1-y^{(i)})\cdot\log(1-\hat{p}^{(i)}) + y^{(i)}\cdot\log(\hat{p}^{(i)})\right]$$

$$= \frac{1}{m}\sum_{i=1}^{m}\underbrace{\left[(1-y^{(i)})\left(-\log(1-\hat{p}^{(i)})\right) + y^{(i)}\left(-\log(\hat{p}^{(i)})\right)\right]}_{\text{perda por elemento}}$$

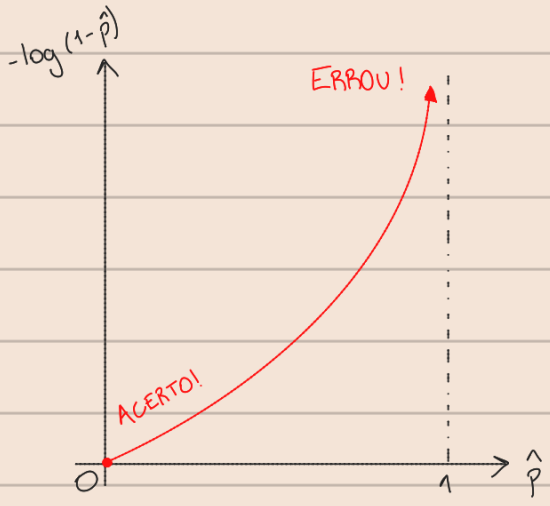
Caso 1: $y^{(i)} = 1$

$$(1-y^{(i)})\left(-\log(1-\hat{p}^{(i)})\right) + y^{(i)}\left(-\log(\hat{p}^{(i)})\right)$$
$$= (1-1)\left(-\log(1-\hat{p}^{(i)})\right) + 1\cdot\left(-\log(\hat{p}^{(i)})\right)$$
$$= -\log(\hat{p}^{(i)})$$

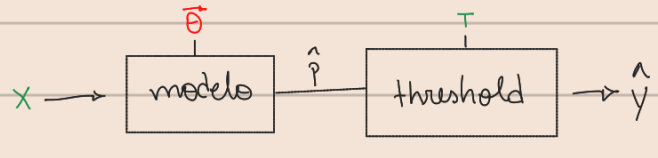


Caso 2: $y^{(i)} = 0$

$$(1-y^{(i)})\left(-\log(1-\hat{p}^{(i)})\right) + y^{(i)}\left(-\log(\hat{p}^{(i)})\right)$$
$$= (1-0)\left(-\log(1-\hat{p}^{(i)})\right) + 0\cdot\left(-\log(\hat{p}^{(i)})\right)$$
$$= -\log(1-\hat{p}^{(i)})$$



Regressão Logística



$$\hat{p} = \sigma(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots \theta_n x_n)$$

função sigmoide