CS 182 Lecture 2: ML Basics 6/6

Types of learning: → supervised, → unsupervised → reinforcement

Supervised learning: Given $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, learn $f_\theta(x) \approx y$

learn $p_\theta(y|x)$ → Predicting probabilities often makes more sense than predicting labels. → must be positive, sum to 1.

$p(x, y) = p(x)p(y|x)$ (chain rule)

$$\text{softmax}_k(f_1(x), \ldots, f_n(x)) = \frac{e^{f_k(x)}}{\sum_{i=1}^{n} e^{f_i(x)}} = p(y = k|x)$$

The ML Method [Example in lecture: Logistic Regression]

1. Define your model class (how to represent the "program")
2. Define your loss function (how to determine "better" model)
3. Pick your optimizer (how to find "best" model)
4. Run on big GPU

$(x, y) \sim p(x, y)$ $\quad p(\mathcal{D}) = \prod_i p(x_i, y_i)$ (i.i.d.) $= \prod_i p(x_i)p(y_i|x_i)$

$\log p(\mathcal{D}) = \sum_i \log p(x_i) + \log p_\theta(y_i|x_i) = \sum_i \log p_\theta(y_i|x_i) + \text{const}$

ideal → $\theta^* \leftarrow \arg\max_\theta \sum_i \log p_\theta(y_i|x_i)$ maximum likelihood estim. (MLE)

$\theta^* \leftarrow \arg\min_\theta \left(-\sum_i \log p_\theta(y_i|x_i)\right)$ negative log-likelihood (NLL) ↳ loss func.

cross-entropy: how similar are two distributions p_θ and p? ~ NLL.

$$H(p, p_\theta) = -\sum_y p(y|x_i) \log p_\theta(y|x_i) \approx -\log p_\theta(y_i|x_i)$$

Optimization algo: (one approach)	Gradient descent
1. Find direction v where $\mathcal{L}(\theta)\downarrow$	1. Compute $\nabla_\theta \mathcal{L}(\theta)$
2. $\theta \leftarrow \theta + \alpha v$ ← learning rate	2. $\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}(\theta)$

gradient: $\nabla_\theta \mathcal{L}(\theta) = \begin{bmatrix} \frac{d\mathcal{L}(\theta)}{d\theta_1} \\ \vdots \\ \frac{d\mathcal{L}(\theta)}{d\theta_n} \end{bmatrix}$

logistic Equation: $\frac{1}{1+e^{-\theta^T x}}$ (aka sigmoid)

can examine over/under-fitting →
- Risk $= \mathbb{E}_{x\sim p(x), y\sim p(y|x)}[\mathcal{L}(x, y, \theta)]$ (Theoretical)
- Empirical Risk $= \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(x_i, y_i, \theta) \approx$ Risk (Practical)