Conditional models

Directly estimate $P(y|x) \sim f(y;\ \theta = g(x))$

Suppose $y \in \mathbb{R}$ eg: regression tasks.

$$P(y|x) \sim \mathcal{N}(y;\ \mu_x = \underbrace{g(x)}\ ;\ \sigma^2 = \text{constant})$$

linear function

$g(x) = W^T x$

Neural network

μ

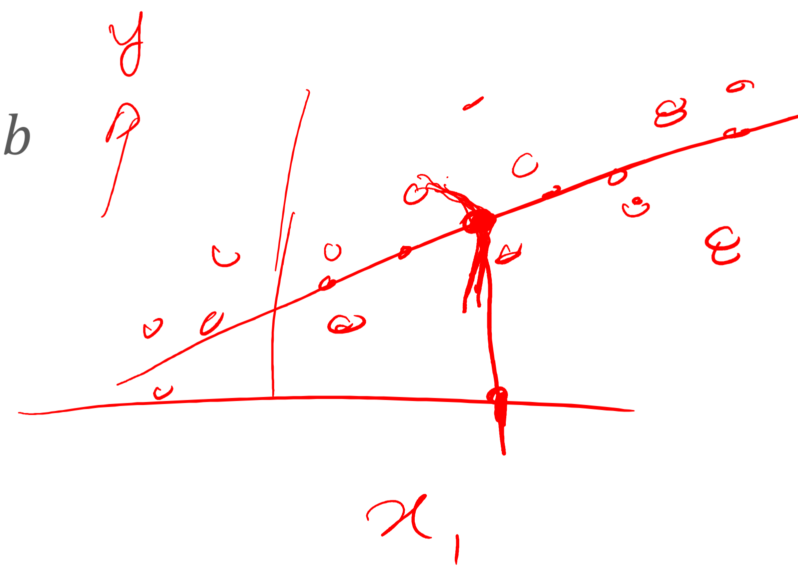
$g(x) \equiv \boxed{NN}$

x

Conditional Probabilistic Approach:

- We will model the conditional distribution: P(y | x): Instead of a single value, we predict a distribution over values to reflect uncertainty.
- Example: regression models.
    - $P(y|x) \sim N(y; \mu_x, \sigma)$
    - $\mu_x = \underbrace{w^T.x + b}_{g(x)} = w_1x_1 + \cdots + w_dx_d + b$
        - Mean prediction is a linear function of x.
    - $\sigma = independent\ of\ x.$
- 1-d diagram

$d = 1$



Estimating parameters using MLE

Earlier (first 2 classes) - we used error minimization as ~~and~~ our training objective:

$$\min_{w,b} L(w,b) = \sum_{i=1}^{N} \left(y_i - (w^T x^i - b)\right)^2$$

contrast above with the MLE approach:

Maximize the probability of the training dataset

$$w^*, b^* = \max_{w,b} \sum_{i=1}^{N} \log \underbrace{\Pr(y^i \mid x^i, w, b)}_{}$$

$$= \max_{w,b} \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi}\,\sigma} e^{\frac{-\left(y^i - (w^T x^i + b)\right)^2}{2\sigma^2}} \iff \min_{w,b} \sum_{i=1}^{N} \left(y^i - (w^T x^i + b)\right)^2$$

Classification (Linear) --- Logistic Classifier

- y is binary, $x \in R^d$
- Conditional Probabilistic Approach:
  - We will model the conditional distribution: P(y | x)
  - $P(y|x) \sim Bernoulli(y; \theta_x)$
- How to obtain Bernoulli parameter (has to be between 0 and 1) from x?
  - Compute a linear function x: $g(x) = w.x + b, \quad g(x) \in [-\infty, \infty]$
  - Use a sigmoid function to squash g(x) between 0 and 1.

Extending Logistic Regression to Multi-class

- A class label y can take one ~~of k~~ possible discrete values.

- Pr(y|x) ~ Multinomial distribution with k parameters each of which is a function of x

- $\Pr(y|x) \sim Mult(y; \{\theta_1(x), \ldots, \theta_k(x)\})$

$p_1(x)$ $p_k(x)$

; $\theta_r(x) \geq 0 \mid \sum_{r=1}^{k} \theta_r(x) = 1$

$\vec{\theta}(x)$ is a simplex.

$(w^1, b^1), \ldots (w^k, b^k)$ k sets of parameters.

Softmax

$$\boxed{\theta_r(x) = \frac{e^{w^r \cdot x + b^r}}{\sum_{r'=1}^{k} e^{w^{r'} \cdot x + b^{r'}}}}$$

normalizer.

Estimating parameters

(Dropping b for convenience)

Again apply MLE.

$$\max_{w^1 \dots w^k} \sum_{i=1}^{N} \log P(y^i \mid x^i, w^1 \dots w^k)$$

$$\max_{w_1 \dots w_k} \sum_{i=1}^{N} \sum_{y=1}^{K} \delta(y == y_i) \log \theta_y(x)$$

$$= \max_{w_1 \dots w_k} \sum_{i=1}^{N} \log \frac{e^{w^{y_i} \cdot x^i}}{\sum_y^{K} e^{w^y \cdot x^i}}$$

$$\boxed{\max_{w^1 \dots w_k} \sum_{i=1}^{N} w^{y_i} \cdot x^i - \log \sum_y e^{w^y \cdot x^i}}$$

Reference:

Earlier error-based objective for logistic regression classifier.

$$L(w_1 \dots w_k) = \boxed{\min_{w^1 \dots w^k} \sum_{i=1}^{N} -w^{y_i} \cdot x^i + \sum_{i=1}^{N} \log\left(\sum_y e^{w^y x^i}\right)}$$

Training NN

Momentum-based opt method

→ identifying convex fns.

→ solve problem similar to sample questions.

→ Q5: logistic regression question