ML & Opt Lecture 5

- Parametrized ML models (GLMs)
- Principle of maximum likelihood estimation (MLE)
- Decomposing risk of ML models into:
  approximal error, generalization, optimization error
  error

Last class

$p(y|x)$ — captures everything about the dependence b/w y & x

$\mathbb{E}[y|x]$ — Bayes optimal (in least squares loss) point estimate of y given x

Issues: either of these could be arbitrarily complicated

To deal with this, we use parametrized models, i.e.

$$p(y|x) = f_\theta(x) \quad \text{where } \theta \in \mathbb{R}^p \text{ is a parameter vector}$$

$$= f(x; \theta)$$

Ex: OLS

$$y = \beta^T x + \varepsilon \qquad \text{where} \quad \varepsilon \sim \mathcal{N}(0, \tfrac{1}{n})$$

$$\text{so} \qquad y|x \sim \mathcal{N}(\beta^T x, \tfrac{1}{n})$$

$$\text{or} \qquad p(y|x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y-\beta^T x)^2}{2\sigma^2}\right)$$

$$\text{where } \sigma = \frac{1}{\sqrt{n}}$$

$\mathbb{E}[y|x] = \beta^T x$ is our regression function

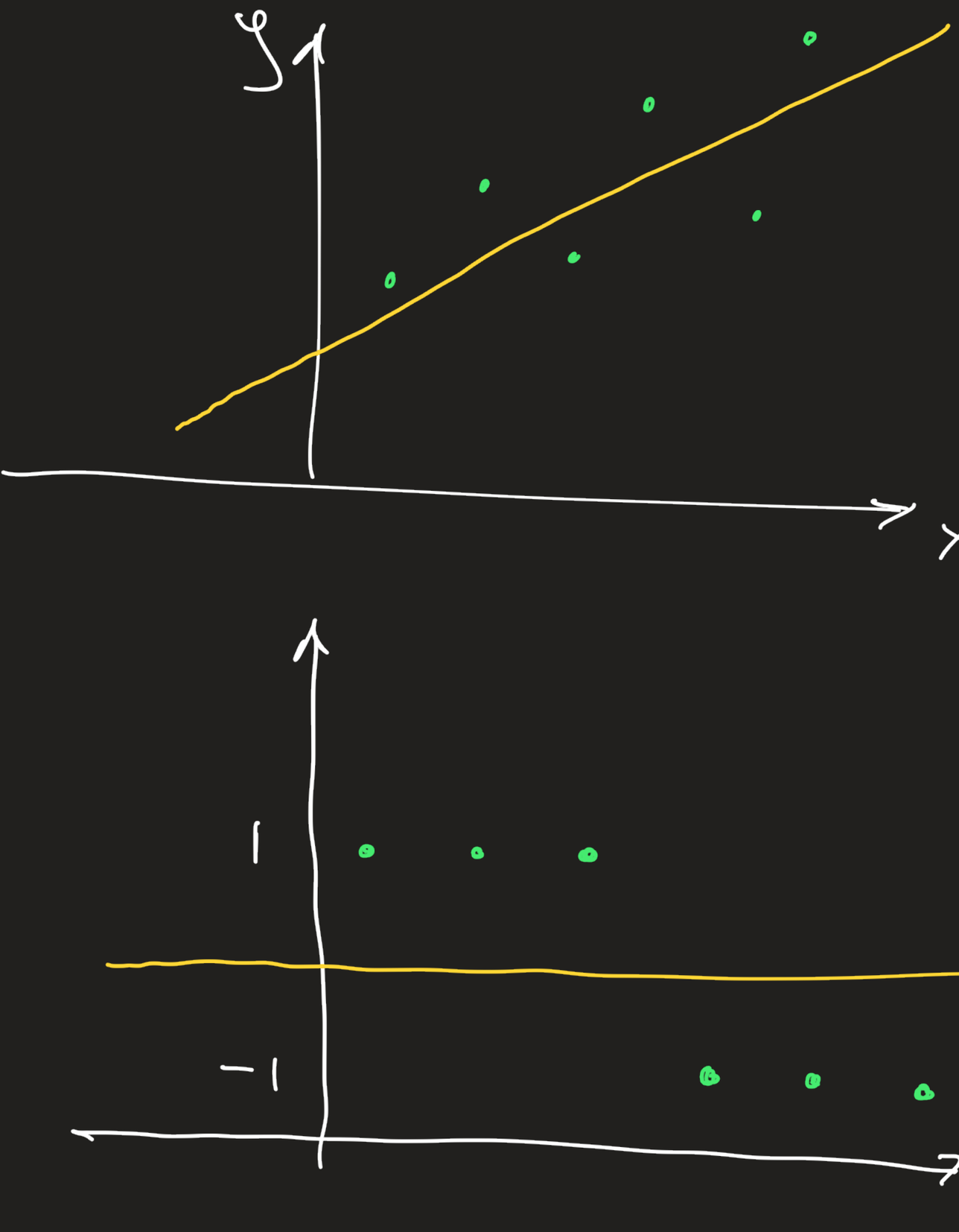
and we estimate $\hat{\beta} = X^+ y$ from our training data

$$X = \underbrace{\begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}}_{d} \qquad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

and showed last time that

$$\mathbb{E}\left[\|\hat{\beta} - \beta\|_2^2 \,|\, X\right] = O\left(\frac{d}{n}\right)$$

OLS is appropriate when data looks like, e.g.



not appropriate for OLS when $y \in \{-1, 1\}$

What about models for different types of data?

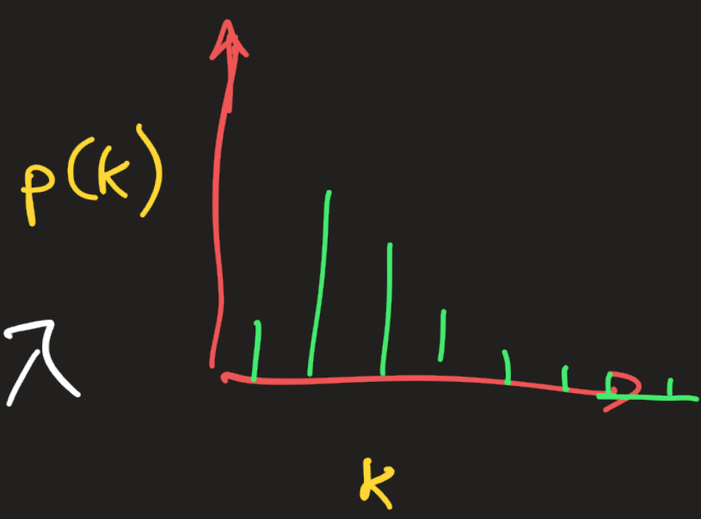
- meaning, e.g. y is not a continuous r.v.,
  y has limited range,
  $y - \mathbb{E}[y|x]$ is not independent of x

Exs

- Poisson / "shot noise" model

corresponds to Poisson(λ) is a r.v. that takes on values in $\mathbb{N}_0 = \{0, 1, 2, \dots\}$ with pmf given by

$$p(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \dots$$



where λ is called the rate, b/c $\mathbb{E}X = \lambda$

We use this in ML for predicting count data

$$y|x \sim \text{Poisson}(\lambda(x))$$

this means

$$\mathbb{E}[y|x] = \lambda(x)$$

Notice since $\lambda(x)$ is a rate, it must satisfy $\lambda(x) > 0$.

We impose this by taking

$$\lambda(x) = e^{\theta^T x}$$

so

$$\mathbb{E}[y|x] = e^{\theta^T x}$$

(n.b. we always assume the covariates x include the constant feature 1, i.e. $x = [x_1, \dots, x_d, 1]$)

- Bernoulli "noise" model

$$y|x \sim \text{Bern}(p(x))$$

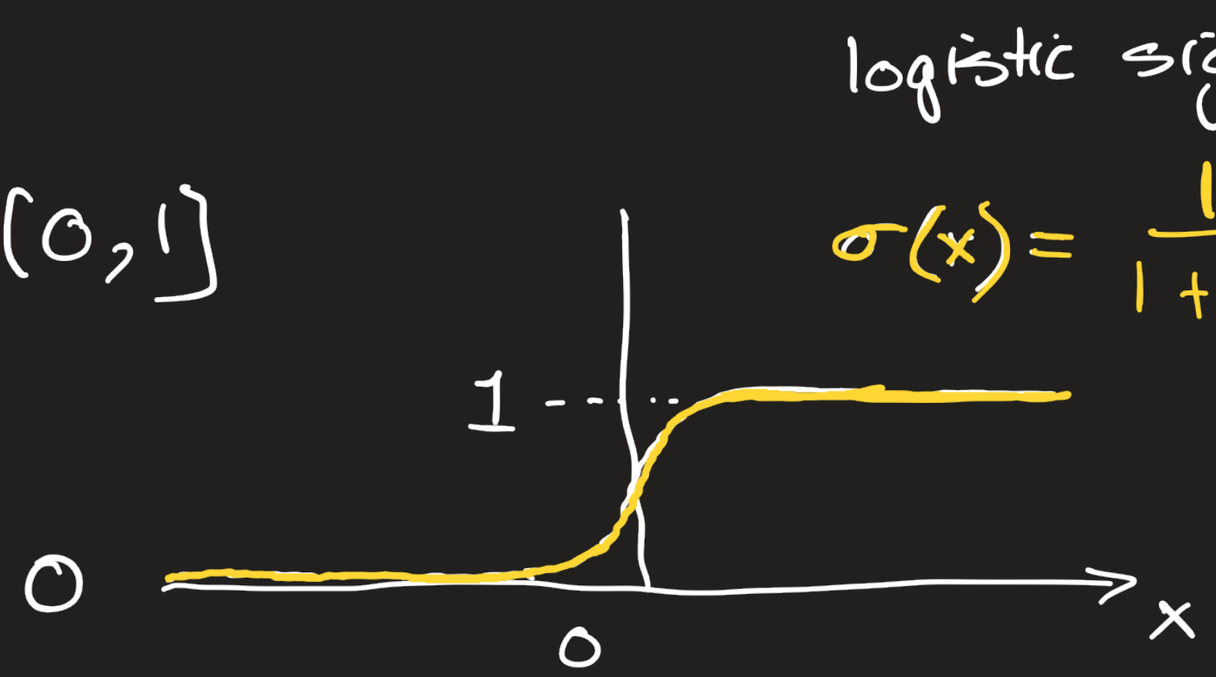
useful for classification (e.g. spam detection)

$$\mathbb{E}[y|x] = p(x)$$

and we require $p(x) \in [0,1]$

logistic sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



$$p(x) = \sigma(\theta^T x)$$

so

$$y|x \sim \text{Bern}(\sigma(\theta^T x))$$

$$\mathbb{E}[y|x] = \underline{\sigma(\theta^T x)}$$

Very used for binary classification

- Categorical regression (multiclass classification) model

$$y|x \sim \text{Categorical}(p_1(x), \cdots, p_K(x))$$

where K is the number of categories and

$$p(y = e_i = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \leftarrow i\text{th} \mid x) = p_i(x)$$

Note that we need $p_i(x) \geq 0$ and $\sum_{i=1}^{K} p_i(x) = 1$.

We do this by taking

$$p_i(x) = \frac{e^{\theta_i^T x}}{Z_\theta(x)}$$

where $Z_\theta(x)$ is the "normalizing constant"

$$\Rightarrow Z_\theta(x) = \sum_{i=1}^{k} e^{\theta_i^T x}$$

Introduce some notation:

$$\Theta = \begin{bmatrix} \theta_1^T \\ \vdots \\ \theta_k^T \end{bmatrix} \in \mathbb{R}^{k \times d}, \quad \underline{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^k$$

and if $x \in \mathbb{R}^k$ then interpret $e^x = \begin{bmatrix} e^{x_1} \\ \vdots \\ e^{x_k} \end{bmatrix}$

Then we have

$$p_\theta(x) = \begin{bmatrix} p_1(x) \\ \vdots \\ p_k(x) \end{bmatrix} = \begin{bmatrix} e^{\theta_1^T x} \\ \vdots \\ e^{\theta_k^T x} \end{bmatrix} \cdot \frac{1}{Z_\theta(x)} = \frac{e^{\Theta x}}{\underline{1}^T e^{\Theta x}}$$

and

$$y|x \sim \text{Categorical}(p_\theta(x))$$

$$\mathbb{E}(y|x) = [p_\theta(x)]_1 e_1 + \cdots + [p_\theta(x)]_k e_k = \begin{bmatrix} p_\theta(x)_1 \\ \vdots \\ p_\theta(x)_k \end{bmatrix} = p_\theta(x)$$

How to interpret θ ?

$$\frac{p(y=e_i|x)}{p(y=e_j|x)} = \frac{e^{\theta_i^T x}/Z_\theta(x)}{e^{\theta_j^T x}/Z_\theta(x)} = \frac{e^{\theta_i^T x}}{e^{\theta_j^T x}}$$

$$= e^{(\theta_i - \theta_j)^T x}$$

In general we can fit an entire class of models, called Generalized Linear Models (GLMs) by choosing

$$y|x \sim \text{Exp}(p(x))$$

where Exp is a particular kind of exponential family distribution and $p(x)$ is a parameter vector.

For a given Exp

$$\mathbb{E}[y|x] = g^{-1}(\theta^T x)$$ where g is called the link function

(e.g. for the Bernoulli model

$$\mathbb{E}[y|x] = \sigma(\theta^T x) \text{ so } g^{-1} = \sigma \Rightarrow g(x) = \ln\left(\frac{x}{1-x}\right)$$)

Maximum Likelihood Estimation

We fixed a model that we expect to work well for our problem. How do we estimate the parameters of that model?

parametrized

$$y|x \sim p_\theta(x)$$

and given $\{(x_i, y_i)\}_{i=1}^n$

To estimate the optimal parameters θ we use the maximum likelihood principle

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} p_\theta(y_i|x_i)$$

This process is called Maximum likelihood estimation.

Notice

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} p_\theta(y_i|x_i)$$

$$= \underset{\theta}{\operatorname{argmax}} \left(\prod_{i=1}^{n} p_\theta(y_i|x_i) \right)^{1/n}$$

$$= \underset{\theta}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(y_i|x_i)$$

$$= \underset{\theta}{\operatorname{argmin}} \underbrace{\frac{1}{n} \sum_{i=1}^{n} -\log p_\theta(y_i|x_i)}$$

called negative log likelihood of θ (NLL)

Ex: MLE for Gaussian model

$$y = \beta^T x + \varepsilon \quad \text{where} \quad \varepsilon \sim N(0, \sigma^2)$$

or equivalently

$$P_\beta(y_i | x_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(- \frac{(y_i - \beta^T x_i)^2}{2\sigma^2} \right)$$

use MLE to estimate β:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} -\log P_\beta(y_i | x_i)$$

$$= \underset{\beta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} - \left[\log\left(\frac{1}{\sqrt{2\pi}\sigma} \right) - \frac{(y_i - \beta^T x_i)^2}{2\sigma^2} \right]$$

$$= \underset{\beta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \beta^T x_i)^2}{2\sigma^2} + \log\left(\sqrt{2\pi}\sigma \right)$$

$$\Rightarrow \hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{n2\sigma^2} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2$$

because σ is a constant

$$= \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2$$

$$= \underset{\beta}{\operatorname{argmin}} \| X\beta - y \|_2^2$$

$$= X^{\dagger} y = \beta_{OLS}$$

so MLE for Gaussian model leads to OLS