ML and Optim Lecture 5

Last time discussed three models $y|x$:

- Gaussian noise model

  Regression $\quad y|x \sim \mathcal{N}(\beta^T x + \beta_0, \sigma^2)$

  $\Rightarrow \mathbb{E}[y|x] = \beta^T x + \beta_0$

- Bernoulli noise model

  Binary Classification $\{0,1\}$ $\quad y|x \sim \text{Bernoulli}(\sigma(\beta^T x + \beta_0))$

  $\mathbb{E}[y|x] = \sigma(\beta^T x + \beta_0)$

- Categorical noise model

  Multiclass Classification $\{1,\dots,K\}$ $\quad y|x \sim \text{Categorical}(p_\theta(x))$

  $p_\theta(x) = \text{softmax}(\theta x) \in \mathbb{R}^K$

  $= \frac{e^{\theta x}}{1^T e^{\theta x}}$

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

Interpretation of Logits

Binary classification case

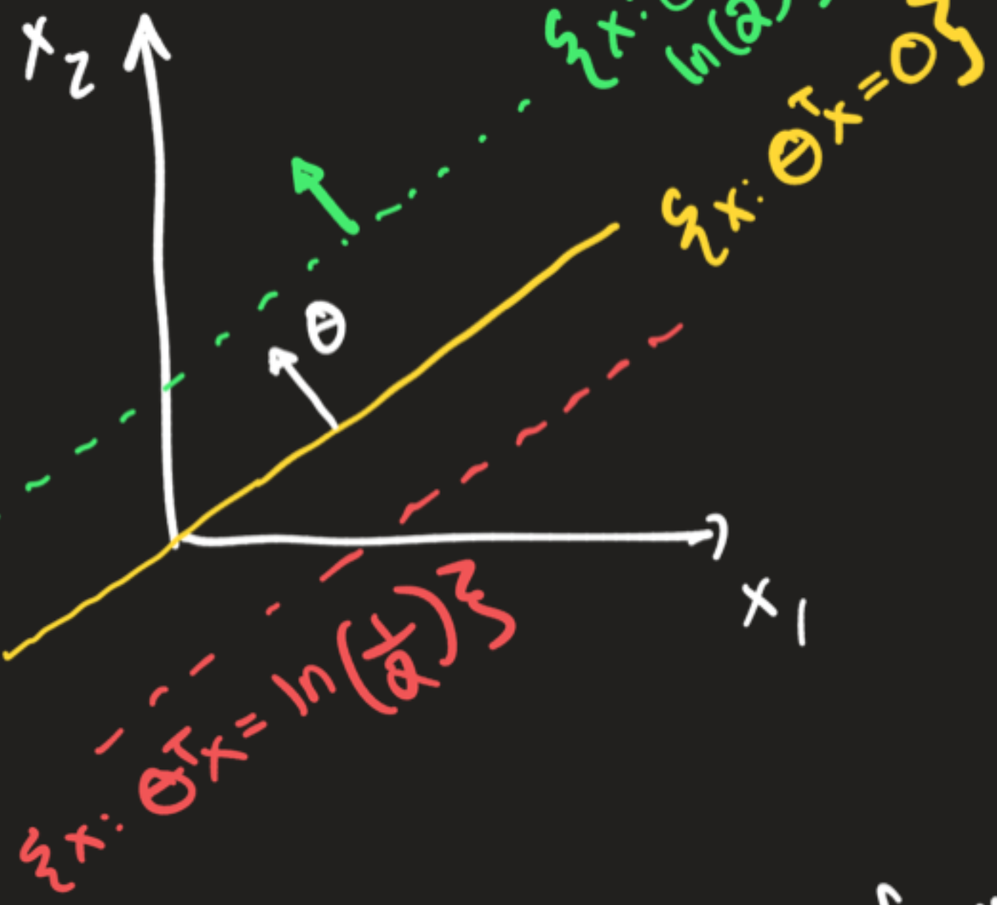
$$p_\theta(1|x) = \sigma(\theta^T x)$$
$$p_\theta(0|x) = 1 - \sigma(\theta^T x)$$

$$\Rightarrow \frac{p(y=1|x)}{p(y=0|x)} = \frac{\sigma(\theta^T x)}{1-\sigma(\theta^T x)}$$

$$= \frac{\frac{1}{1+e^{-\theta^T x}}}{\left(1 - \frac{1}{1+e^{-\theta^T x}}\right)} = \frac{\frac{1}{1+e^{-\theta^T x}}}{\left(\frac{1+e^{-\theta^T x}}{1+e^{-\theta^T x}} - \frac{1}{1+e^{-\theta^T x}}\right)}$$

$$= \frac{\frac{1}{1+e^{-\theta^T x}}}{\left(\frac{e^{\theta^T x}}{1+e^{-\theta^T x}}\right)} = e^{\theta^T x}$$

$$\Rightarrow \log\left[\frac{p(y=1|x)}{p(y=0|x)}\right] = \theta^T x$$



Aside

$\{x: \theta^T x = 0\}$ is a line, geometrically, with normal vector θ

In old case, if $x = \{y, z\}$

then $\theta^T x = 0 \Leftrightarrow \theta_1 y + \theta_2 z = 0$

$\Leftrightarrow y = \frac{\theta_2}{\theta_1} z$

Geometric interpretation of the logits here:

$\theta^T x = c$ is exactly the set of inputs where

$$\frac{p(y=1|x)}{p(y=0|x)} = e^c$$

Multiclass classification

Consider log odds of class i vs j:

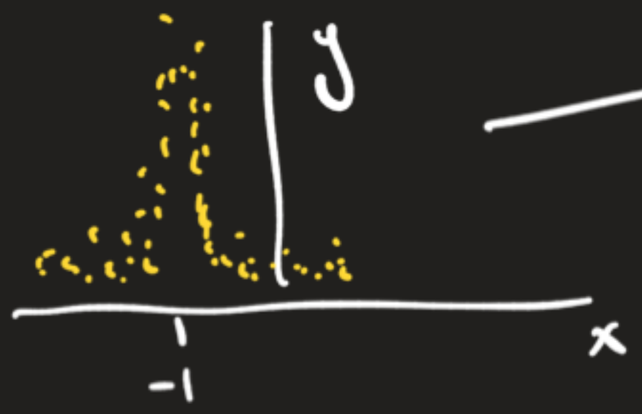
$$\log\left[\frac{p(y=e_i \mid x)}{p(y=e_j \mid x)}\right] = \log\left[\frac{e^{\theta_i^T x}/z_\theta(x)}{e^{\theta_j^T x}/z_\theta(x)}\right]$$

$$\Rightarrow \log\left[\frac{p(y=e_i \mid x)}{p(y=e_j \mid x)}\right] = (\theta_i - \theta_j)^T x = \theta_i^T x - \theta_j^T x$$

Geometric interpretation:

x should be labeled i rather than j if $\theta_i^T x \gg \theta_j^T x$

Maximum Likelihood Estimation (MLE)



Given observations y that came from a Gaussian, we fit a model

$$y \sim \mathcal{N}(-1, \sigma^2)$$

because the data is best explained w/ this choice

Q: how to fit model parameters given data (x, y)?

A: choose θ so the joint probability of the observed data under this model is maximized

(maximum likelihood principle)

MLE

Assume our i.i.d.! training data $\{(x_i, y_i)\}_{i=1}^n$ follows a parameterized model

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

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

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ p_\theta(y_1, \dots, y_n | x_1, \dots, x_n)$$

likelihood of θ under these observations

because data is i.i.d.

$$= \underset{\theta}{\operatorname{argmax}}\ p_\theta(y_1|x_1)\, p_\theta(y_2|x_2) \cdots p_\theta(y_n|x_n)$$

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

Numerically dealing w/ products is unwieldy, so use

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

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

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

$$= \operatorname*{argmin}_{\theta} \underbrace{\frac{1}{n} \sum_{i=1}^{n} -\log p_\theta(y_i|x_i)}_{:= \mathcal{L}(\theta)}$$

Called the negative log-likelihood of θ (NLL)

Aside: If f is a strictly increasing function ($f(x) > f(y)$ when $x > y$)

Then

$$\underset{\theta}{\operatorname{argmax}}\ J(\theta) = \underset{\theta}{\operatorname{argmax}}\ f(J(\theta))$$

Issues that arise in MLE (minimizing NLL)

Q: does a minimizer exist (Q: for the models we gave)

Q: is it unique (maybe)

Q: can we find the minimizer efficient
(Q: for the models we gave)

- Ex: MLE for Gaussian model

$$y = \beta^T x + \beta_0 + \varepsilon \quad \text{where} \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

or equiv

$$p_{\beta,\sigma^2}(y_i|x_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \beta^T x_i - \beta_0)^2}{2\sigma^2}\right)$$

so

$$\mathcal{L}(\beta, \sigma^2) = \frac{1}{n} \sum_{i=1}^{n} -\log p_{\beta,\sigma^2}(y_i|x_i)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left[\log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{(y_i - \beta^T x_i - \beta_0)^2}{2\sigma^2}\right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left[\log(\sqrt{2\pi}\sigma) + \frac{1}{2\sigma^2}(y_i - \beta^T x_i - \beta_0)^2\right]$$

We could learn β, β_0, σ^2 by minimizing this NLL

In practice σ^2 is often ignored as a nuisance parameter, so instead we choose β by minimizing

$$\mathcal{L}(\beta) = \frac{1}{2\sigma^2 n}\sum_{i=1}^{n}(y_i - x_i^T\beta - \beta_0)^2$$

$$\sim \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\beta - \beta_0)^2$$

← ERM w/ $\ell(v,v) = (v-v)^2$

So "MLE for linear Gaussian noise model is OLS"!

MLE for Bernoulli noise model

$$y|x \sim \text{Bernoulli}(\sigma(\theta^T x)) \quad \text{(for } \pm 1 \text{ labels!)}$$

$$\mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} -\log p_\theta(y_i|x_i)$$

Check the identity $1-\sigma(t) = \sigma(-t)$

Notice $p_\theta(1|x_i) = \sigma(\theta^T x_i)$

$p_\theta(-1|x_i) = 1-\sigma(\theta^T x_i) = \sigma(-\theta^T x_i)$

These two equations imply that

$$p_\theta(y_i|x_i) = \sigma(y_i \theta^T x_i)$$

$$\Rightarrow \mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} -\log \sigma(y_i \theta^T x_i) = \frac{1}{n} \sum_{i=1}^{n} -\log \left[\frac{1}{1+e^{-y_i \theta^T x_i}} \right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} \log(1+e^{-y_i \theta^T x_i})$$ ← ERM w/ $\ell(v,v) = \log(1+e^{-uv})$

$$\hat{\theta} = \operatorname*{argmin}_{\theta} \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + e^{-y_i \theta^T x_i}\right)$$

↑ this is logistic regression (w/ ± 1 labels)

MLE for Categorical Model

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

where $$p_i(x) = \frac{e^{\theta_i^T x}}{1^T e^{\theta^T x}} = \text{softmax}(\theta x)_i$$

$$\theta = \begin{bmatrix} \theta_1^T \\ \vdots \\ \theta_K^T \end{bmatrix} \in \mathbb{R}^{K \times d}$$

Notice

$$\mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} -\log p_\theta(y_i|x_i)$$

where y_i is a one-hot encoding of the class of the i^{th} example

e.g. if example i is class 2, then $y_i = e_2$

So

$$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} -\log p_\theta(y_i|x_i)$$

$$= \frac{1}{n}\sum_{i=1}^{n} -\log\left[y_i^T \,\text{softmax}(\theta x_i)\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} -y_i^T \log\left(\text{softmax}(\theta x_i)\right)$$

apply log entrywise to its vector argument

multiclass logistic regression

$$= \frac{1}{n}\sum_{i=1}^{n} \ell\left(\text{softmax}(\theta x_i), y_i\right)$$

↑ what is this loss? either KL-div or cross-entropy