Applied Machine Learning

Regression: Generalized Linear Models

Regression: Generalized Linear Models

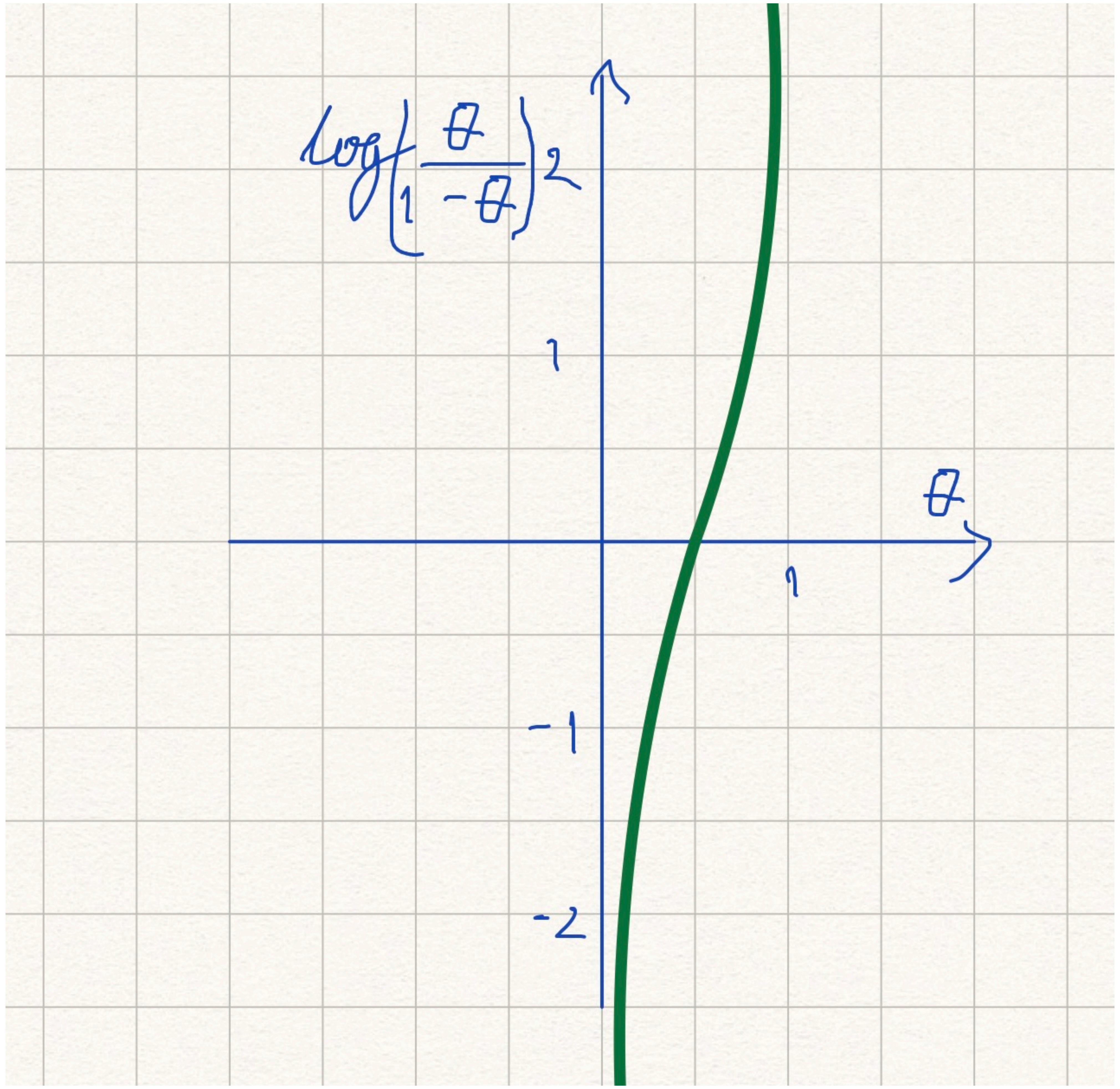
- Generalized Linear Models
- Logistic Regression
- Two more examples
- Performance of GLMs

Generalized Linear Models

- Regression predicts real values
  - $y = \mathbf{x}^\top \beta + \xi$
  - $\xi \sim \mathcal{N}(0, \sigma_x i^2)$
- Generalized Linear Models extend regression to predict probabilities
  - $Y \sim \mathcal{N}(\mathbf{x}^\top \beta, \sigma_\xi^2)$
  - Link function g for a distribution with parameter θ
    - $g : \theta \rightarrow \mathbf{x}^\top \beta$
  - performance is evaluated through *deviance*

GLM: Logistic Regression

- 2-class classification $y \in \{0,1\}$
- $p(y \mid \mathbf{x})$
  - Bernoulli random variable with parameter θ
    - $P(y = 1 \mid \mathbf{x}) = \theta, P(y = 0 \mid \mathbf{x}) = \theta - 1$
    - $\log \left[\frac{P(y = 1 \mid \theta)}{P(y = 0 \mid \theta)} \right] = \log \left[\frac{\theta}{1 - \theta} \right]$
- link function
  - $g : \theta \to \mathbf{x}^\top \beta \mid \theta \in [0,1]$



GLM: Logistic Regression

- Obtaining $p(y \mid \mathbf{x})$

$$
\begin{aligned}
g(\theta) &= \log\left[\frac{P(y=1 \mid \theta)}{P(y=0 \mid \theta)}\right] \\
&= \log\left[\frac{\theta}{1-\theta}\right] = \mathbf{x}^\top \beta \\
\frac{\theta}{1-\theta} &= e^{\mathbf{x}^\top \beta}
\end{aligned}
$$

Link function

-

- $$P(y=1 \mid \mathbf{x}, \beta) = \frac{e^{\mathbf{x}^\top \beta}}{1+e^{\mathbf{x}^\top \beta}} \qquad P(y=0 \mid \mathbf{x}, \beta) = \frac{1}{1+e^{\mathbf{x}^\top \beta}}$$

GLM: Logistic Regression

- Optimization: from $P(y = 1 \mid \mathbf{x})$
- negative log-likelihood $$-\mathscr{L}(\beta) = -\sum_i \left[\mathbb{I}_{y=1}(y_i)\mathbf{x}_i^\top \beta - \log(1 + e^{\mathbf{x}_i^\top \beta}) \right]$$
- Indicator function $$\mathbb{I}_{y=1}(y_i) = \begin{cases} 1 & \text{when } y_i = 1 \\ 0 & \text{otherwise} \end{cases}$$
- Solve through stochastic gradient descent

GLM: Logistic Regression and SVMs

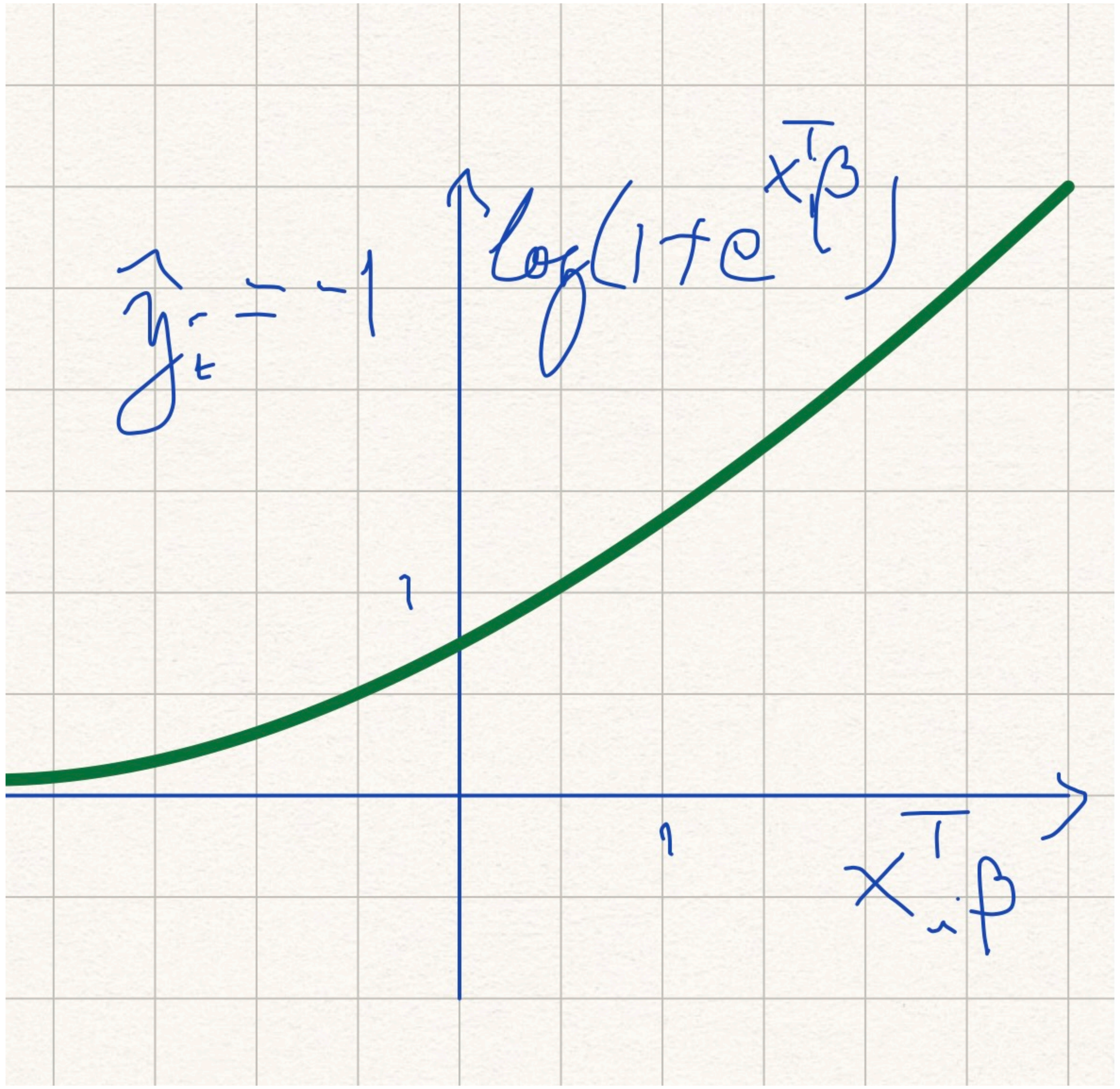
- $\hat{y}_i = 2y_i - 1$ $\quad y \in [0,1] \mapsto \hat{y} \in [-1,1]$
- replace indicator function $\quad \mathbb{I}_{y=1}(y_i) = \dfrac{\hat{y}_i + 1}{2}$
- negative log-likelihood

$$
\begin{aligned}
-\mathscr{L}(\beta) &= -\sum_i \left[\frac{\hat{y}_i + 1}{2}\mathbf{x}_i^\top \beta - \log(1 + e^{\mathbf{x}_i^\top \beta})\right] \\
&= \sum_i \left[\log(e^{-\frac{\hat{y}_i + 1}{2}\mathbf{x}_i^\top \beta}) + \log(1 + e^{\mathbf{x}_i^\top \beta})\right] \\
&= \sum_i \left[\log(e^{-\frac{\hat{y}_i + 1}{2}\mathbf{x}_i^\top \beta} + e^{-\frac{\hat{y}_i - 1}{2}\mathbf{x}_i^\top \beta})\right]
\end{aligned}
$$

GLM: Logistic Regression and SVMs

- negative log-likelihood
$$-\mathscr{L}(\beta) \quad = \quad \sum_i \left[\log(e^{-\frac{\hat{y}_i+1}{2}\mathbf{x}_i^\top\beta} + e^{-\frac{\hat{y}_i-1}{2}\mathbf{x}_i^\top\beta})\right]$$

- for $\hat{y}_i = -1$
$$-\mathscr{L}(\beta) \quad = \quad \sum_i \left[\log(1 + e^{\mathbf{x}_i^\top\beta})\right]$$
    - similar to SVM's hinge loss
- Similar behavior for $\hat{y}_i = 1$



GLM: Multi-class Logistic Regression

- C-class classification $y \in \{0,1,\ldots,C-1\}$
- Discrete probability distribution with parameters $(\theta_0, \theta_1, \ldots, \theta_{C-1})$ so that $\theta_i \in [0,1]$ and $\sum_i \theta_1 = 1$
- Link function: $$g(\theta) \quad = \quad \log\left[\frac{\theta_i}{1-\sum_u \theta_u}\right] \quad = \quad \mathbf{x}^\top \beta$$
- $$\begin{aligned} P(y=0 \mid \mathbf{x}, \beta) &= \frac{e^{\mathbf{x}^\top \beta_0}}{1+\sum_i e^{\mathbf{x}^\top \beta_i}} \\ P(y=1 \mid \mathbf{x}, \beta) &= \frac{e^{\mathbf{x}^\top \beta_1}}{1+\sum_i e^{\mathbf{x}^\top \beta_i}} \\ &\vdots \\ P(y=C-1 \mid \mathbf{x}, \beta) &= \frac{1}{1+\sum_i e^{\mathbf{x}^\top \beta_i}} \end{aligned}$$

GLM: Regression for counting

- Probability of count: $y \in \{0,1,2,\ldots\}$
- Poisson distribution with intensity parameter $\theta > 0$ so that $P(Y = k) = \dfrac{\theta^k e^{-\theta}}{k!}$
- Link function:

$$\begin{aligned} g(\theta) &= \log\left[\theta\right] = \mathbf{x}^\top \beta \\ \theta &= e^{\mathbf{x}^\top \beta_i} \end{aligned}$$

-

$$P(Y = y_i \,|\, \mathbf{x}_i, \beta_i) = \frac{e^{y_i \mathbf{x}_i^\top \beta_i} e^{-e^{\mathbf{x}_i^\top \beta_i}}}{y_i!}$$

$$\begin{aligned} -\mathscr{L}(\beta) &= -\sum_i \log\left(\frac{e^{y_i \mathbf{x}_i^\top \beta_i} e^{-e^{\mathbf{x}_i^\top \beta_i}}}{y_i!}\right) \\ &= -\sum_i \left(y_i \mathbf{x}_i^\top \beta_i - e^{\mathbf{x}_i^\top \beta_i} - \log(y_i!)\right) \\ &= -\sum_i \left(y_i \mathbf{x}_i^\top \beta_i - e^{\mathbf{x}_i^\top \beta_i}\right) + \sum_i \log(y_i!) \end{aligned}$$

GLMs: Performance

- GLMs predict probabilities: $P(y_i \mid \mathbf{x}_i, \hat{\beta})$
  - deviance $= -2 \log P(y_t \mid \mathbf{x}_i, \hat{\beta})$
  - point $\mathbf{x}_i$, true value: y_t
- Linear regression
  - $$-2 \log P(y_t \mid \mathbf{x}_i, \hat{\beta}) = \frac{(\mathbf{x}_i^\top \hat{\beta} - y_t)^2}{\sigma_\xi^2} + K$$

Regression: Generalized Linear Models

- Generalized Linear Models
- Logistic Regression
- Two more examples
- Performance of GLMs

Applied Machine Learning

Regression: Generalized Linear Models