Statistics and Data Science for Engineers
E178 / ME276DS

Logistic regression

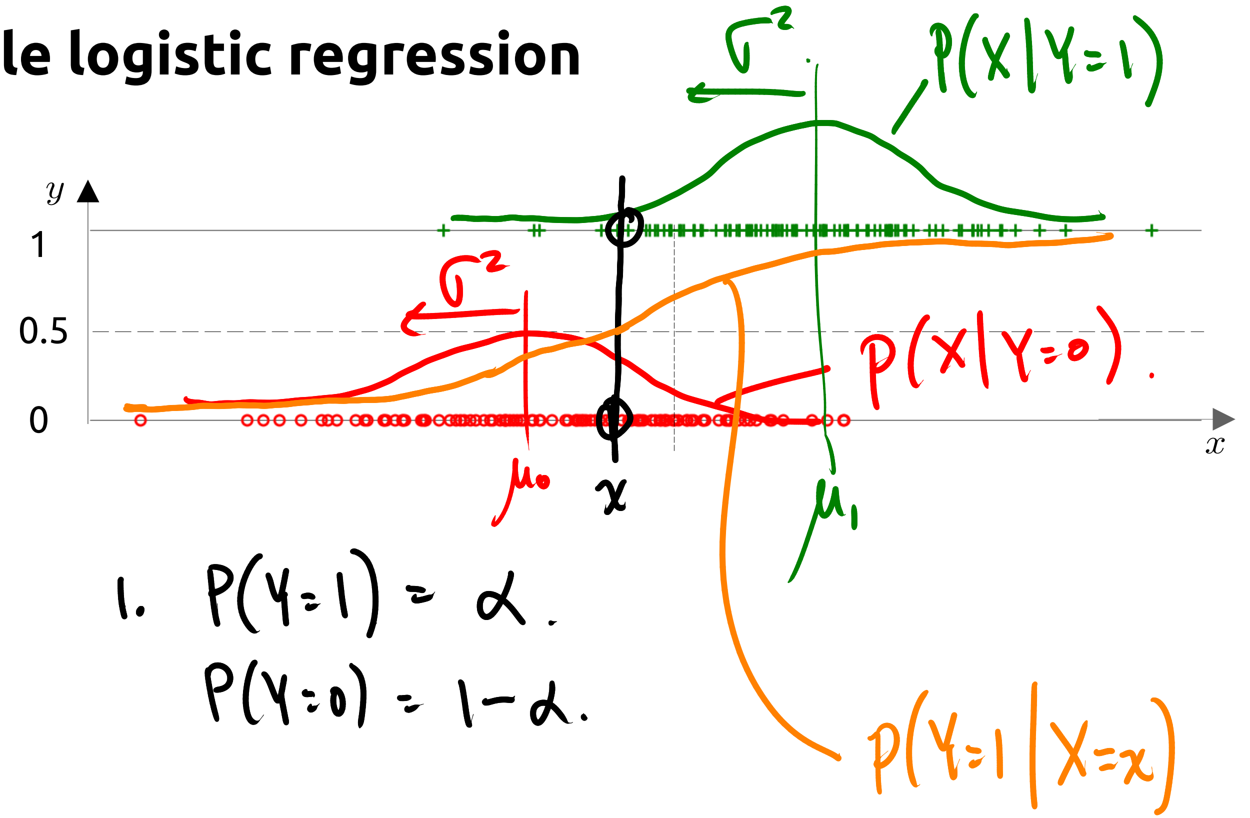
Naïve Bayes: Used assumptions to approximate $p(Y=c)\prod_{d=1}^{D} p(X^d = x^d \mid Y=c)$

Simple logistic regression

Assumptions:

$K=2$.

1. $Y \sim \mathcal{B}(\alpha)$, α is known
2. $D=1$, single input.
3. $X \mid Y=0 \sim \mathcal{N}(\mu_0, \sigma^2)$
4. $X \mid Y=1 \sim \mathcal{N}(\mu_1, \sigma^2)$
5. μ_0, μ_1, σ^2 are known



1. $P(Y=1) = \alpha$.
$P(Y=0) = 1-\alpha$.

$$\hat{y} = h(x) = \underset{c \in \{0,1\}}{\text{argmax}}\, p(Y = c \mid X = x)$$

$$h(x) = \begin{cases} 1 & p(Y=1 \mid X=x) > p(Y=0 \mid X=x) \\ 0 & \text{otherwise} \end{cases}$$

$$= \begin{cases} 1 & \frac{p(Y=1 \mid X=x)}{p(Y=0 \mid X=x)} > 1 \\ 0 & \text{otherwise} \end{cases}$$ ... <u>odds ratio</u> ✓

$$= \begin{cases} 1 & \log\left(\frac{p(Y=1 \mid X=x)}{p(Y=0 \mid X=x)}\right) > 0 \\ 0 & \text{otherwise} \end{cases}$$ ... logs odds ratio ✓

$$\log\left(\frac{p(Y=1|X=x)}{p(Y=0|X=x)}\right) \overset{\text{Bayes' rule}}{=} \log\left(\frac{\frac{p(Y=1)p(X=x|Y=1)}{\cancel{p(X=x)}}}{\frac{p(Y=0)p(X=x|Y=0)}{\cancel{p(X=x)}}}\right)$$

$$= \log\left(\frac{p(Y=1)}{p(Y=0)}\,\frac{p(X=x|Y=1)}{p(X=x|Y=0)}\right)$$

$$= \log\left(\frac{\alpha}{1-\alpha}\,\frac{\cancel{\frac{1}{\sqrt{2\pi\sigma^2}}}\,\exp\left(-\frac{1}{2}\frac{(x-\mu_1)^2}{\sigma^2}\right)}{\cancel{\frac{1}{\sqrt{2\pi\sigma^2}}}\,\exp\left(-\frac{1}{2}\frac{(x-\mu_0)^2}{\sigma^2}\right)}\right)$$

$$\vdots$$

$$\vdots$$

$\cancel{x^2} - \cancel{x^2}$

$$= \log\left(\frac{\alpha}{1-\alpha}\right) - \frac{1}{2\sigma^2}\Big((x-\mu_1)^2 - (x-\mu_0)^2\Big)$$

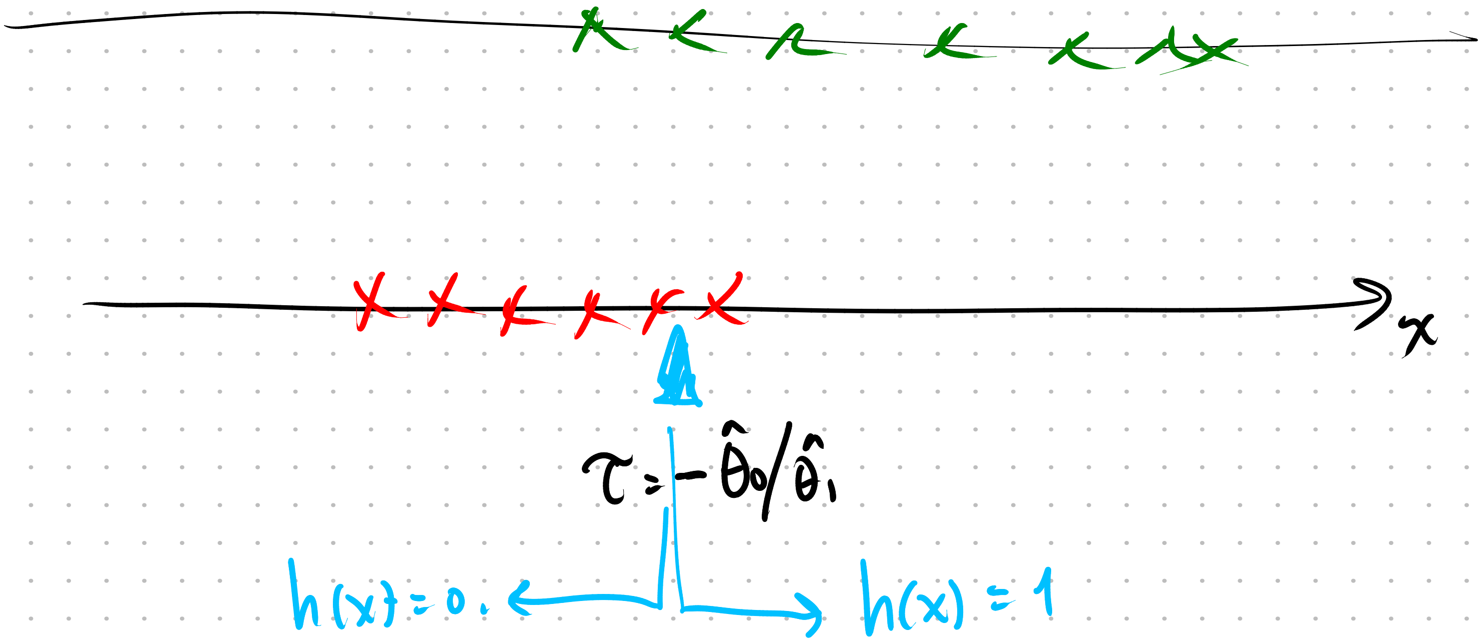
$$= \log\left(\frac{\alpha}{1-\alpha}\right) - \frac{1}{2\sigma^2}\Big(-2(\mu_1-\mu_0)x + (\mu_1^2-\mu_0^2)\Big)$$

$$= \underbrace{\log\left(\frac{\alpha}{1-\alpha}\right) + \frac{\mu_0^2-\mu_1^2}{2\sigma^2}}_{\hat\theta_0} + x\,\underbrace{\frac{\mu_1-\mu_0}{\sigma^2}}_{\hat\theta_1}$$

$$= \hat\theta_0 + x\,\hat\theta_1$$

$$\therefore \qquad h(x) = \begin{cases} 1 & \hat\theta_0 + x\,\hat\theta_1 > 0 \\ 0 & \text{otherwise} \end{cases}$$

$\tau : -\hat\theta_0 / \hat\theta_1$


x
$\tau = -\hat{\theta}_0/\hat{\theta}_1$
$h(x)=0$
$h(x)=1$

Compute the conditional probabilities in terms of $\hat{\theta}_0$ and $\hat{\theta}_1$

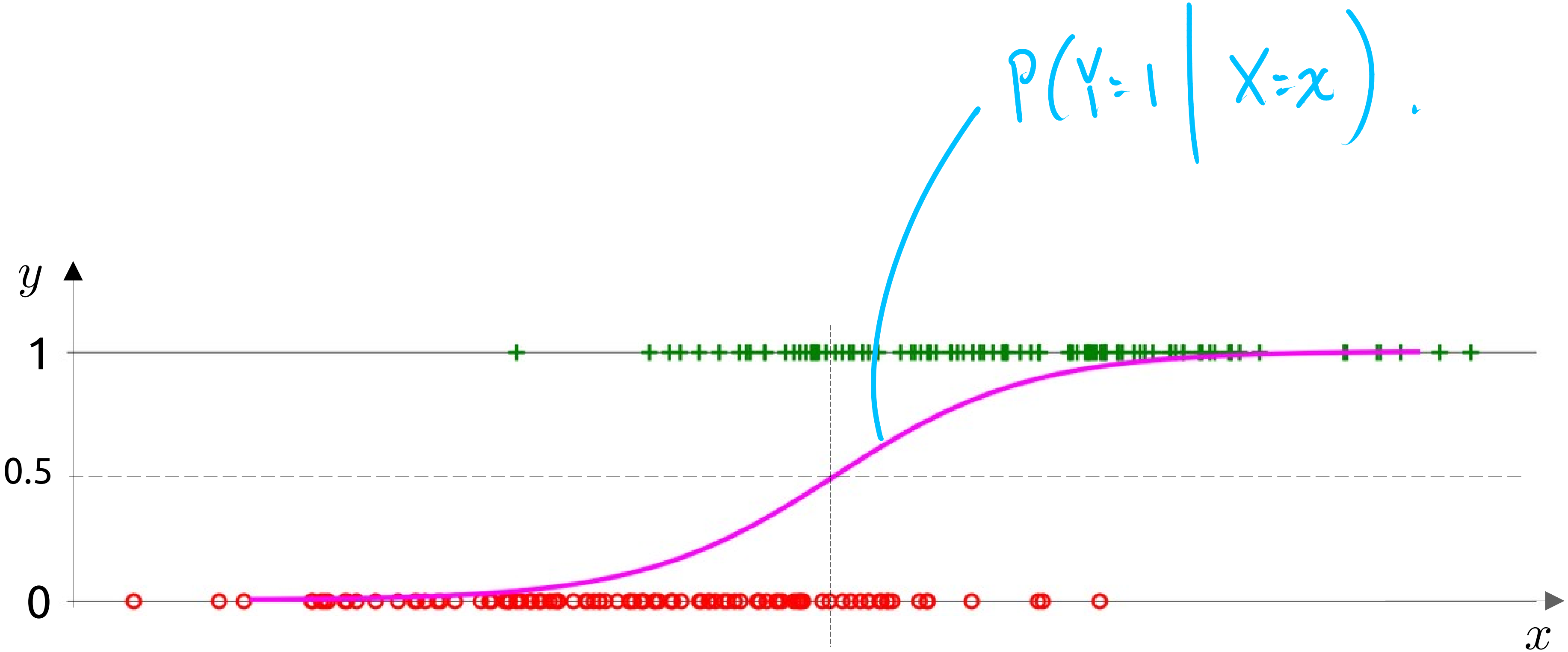
$$\ln\left(\frac{p(Y=1|X=x)}{p(Y=0|X=x)}\right) = \hat{\theta}_0 + \hat{\theta}_1 x$$

$$\Longleftrightarrow \quad \frac{p(Y=1|X=x)}{1-p(Y=1|X=x)} = \exp\left(\hat{\theta}_0 + \hat{\theta}_1 x\right)$$

$$\Longleftrightarrow \quad p(Y=1|X=x) = \frac{1}{1+\exp\left(-(\hat{\theta}_0 + \hat{\theta}_1 x)\right)}$$

Define the ***sigmoid*** function: $\sigma(z) = \dfrac{1}{1+e^{-z}}$ $\quad z = \hat{\theta}_0 + \hat{\theta}_1 x.$

Then $\quad p(Y=1|X=x) = \sigma(\hat{\theta}_0 + \hat{\theta}_1 x)$


P(Y=1 | X=x).
y
1
0.5
0
x

Full logistic regression

$x = \begin{bmatrix} x^1 \\ \vdots \\ x^D \end{bmatrix}$ $\underline{\theta_1} = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_D \end{bmatrix}$

Assumptions:

$K = 2$

1. $\boxed{Y \sim \mathcal{B}(\alpha)}$ ~~α is known~~
2. ~~$D=1$~~,
3. ~~$X \mid Y=0 \sim \mathcal{N}(\mu_0, \sigma^2)$~~
4. ~~$X \mid Y=1 \sim \mathcal{N}(\mu_1, \sigma^2)$~~
5. ~~μ_0, μ_1, σ^2 are known~~

New assumption:

$$
\begin{aligned}
p(Y=1|X=x) &= \sigma(\theta_0 + x^1\theta_1 + \ldots + x^D\theta_D) \\
&= \sigma(\theta_0 + x^T\underline{\theta}_1)
\end{aligned}
$$

for some $\theta_0, \theta_1, \ldots \theta_D$

$$P(Y=0|X=x) = 1 - \sigma(\theta_0 + x^T\underline{\theta_1}).$$

MLE solution to full logistic regression

$$\mathcal{L}(\theta_0, \underline{\theta}_1) = \prod_{i=1}^{N} p(\boxed{y_i | x_i}; \theta_0, \underline{\theta}_1)$$

$$= \prod_{\{i: y_i = 1\}} \sigma(\theta_0 + x_i^T \underline{\theta}_1) \prod_{\{i: y_i = 0\}} (1 - \sigma(\theta_0 + x_i^T \underline{\theta}_1))$$

$$(\hat{\theta}_0, \hat{\underline{\theta}}_1) = \underset{\theta_0, \underline{\theta}_1}{\text{argmax}} \left(\prod_{\{i: y_i = 1\}} \sigma\left(\theta_0 + x_i^T \underline{\theta}_1\right) \prod_{\{i: y_i = 0\}} \left(1 - \sigma(\theta_0 + x_i^T \underline{\theta}_1)\right) \right)$$

clever trick

$$= \underset{\theta_0, \underline{\theta}_1}{\text{argmax}} \prod_{i=1}^{N} \sigma\left(\theta_0 + x_i^T \underline{\theta}_1\right)^{y_i} \left(1 - \sigma\left(\theta_0 + x_i^T \underline{\theta}_1\right)\right)^{(1 - y_i)}$$

log.

$$= \underset{\theta_0, \underline{\theta}_1}{\text{argmax}} \sum_{i=1}^{N} \left(y_i \log \underbrace{\sigma\left(\theta_0 + x_i^T \underline{\theta}_1\right)}_{\text{estimated } P(Y=1 \mid X=x).} + (1 - y_i) \log \left(1 - \sigma\left(\theta_0 + x_i^T \underline{\theta}_1\right)\right) \right)$$

x(-1)

True output

$$= \underset{\theta_0, \underline{\theta}_1}{\text{argmin}} \sum_{i=1}^{N} \underbrace{\text{CE}(y_i, \hat{p}_i)}_{\text{cross entropy.}}$$

where $\hat{p}_i = \sigma(\theta_0 + x_i^T \underline{\theta}_1) \approx P(Y=1 \mid X=x).$

$$\rightarrow \text{CE}(y, p) = -y \log(p) - (1 - y) \log(1 - p)$$

Cross entropy loss

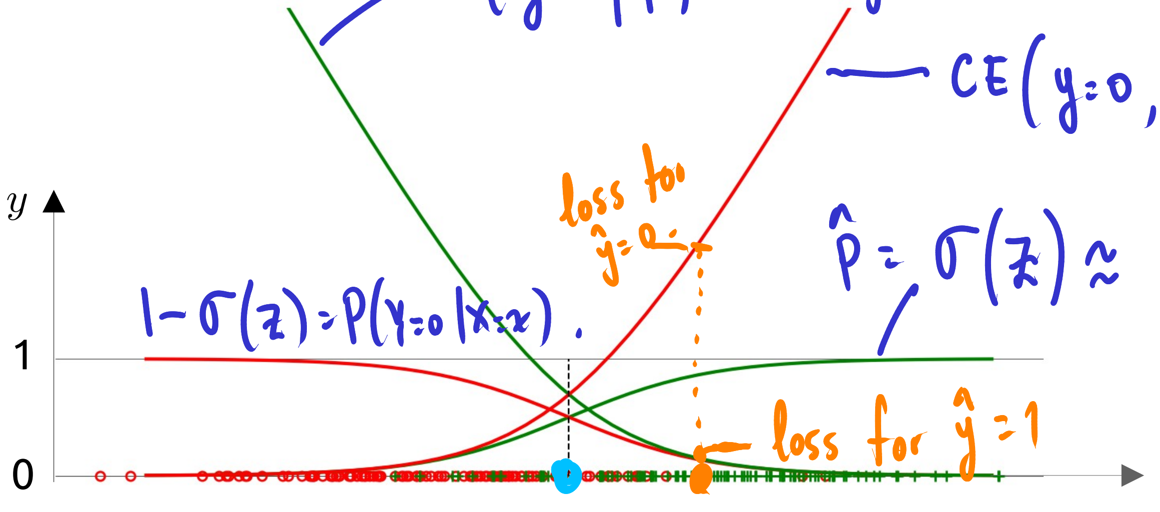
$\{0,1\}$ $\quad$ $(0,1)$

$$\mathsf{CE}(\overset{\{0,1\}}{y},\overset{(0,1)}{p}) = -y\,\mathsf{log}(p) - (1-y)\,\mathsf{log}(1-p) = \begin{cases} -\mathsf{log}(p) & y=1 \\ -\mathsf{log}(1-p) & y=0 \end{cases}$$

$CE(y=1 \mid \hat{p}) = -\log(\hat{p})$

$\hat{p} = \sigma(z)$

$CE(y=0, \hat{p}) = -\log(1-\hat{p})$



loss for $\hat{y}=0$

$\hat{p} = \sigma(z) \approx P(Y=1 \mid X=x)$

$1-\sigma(z) = P(Y=0 \mid X=x)$

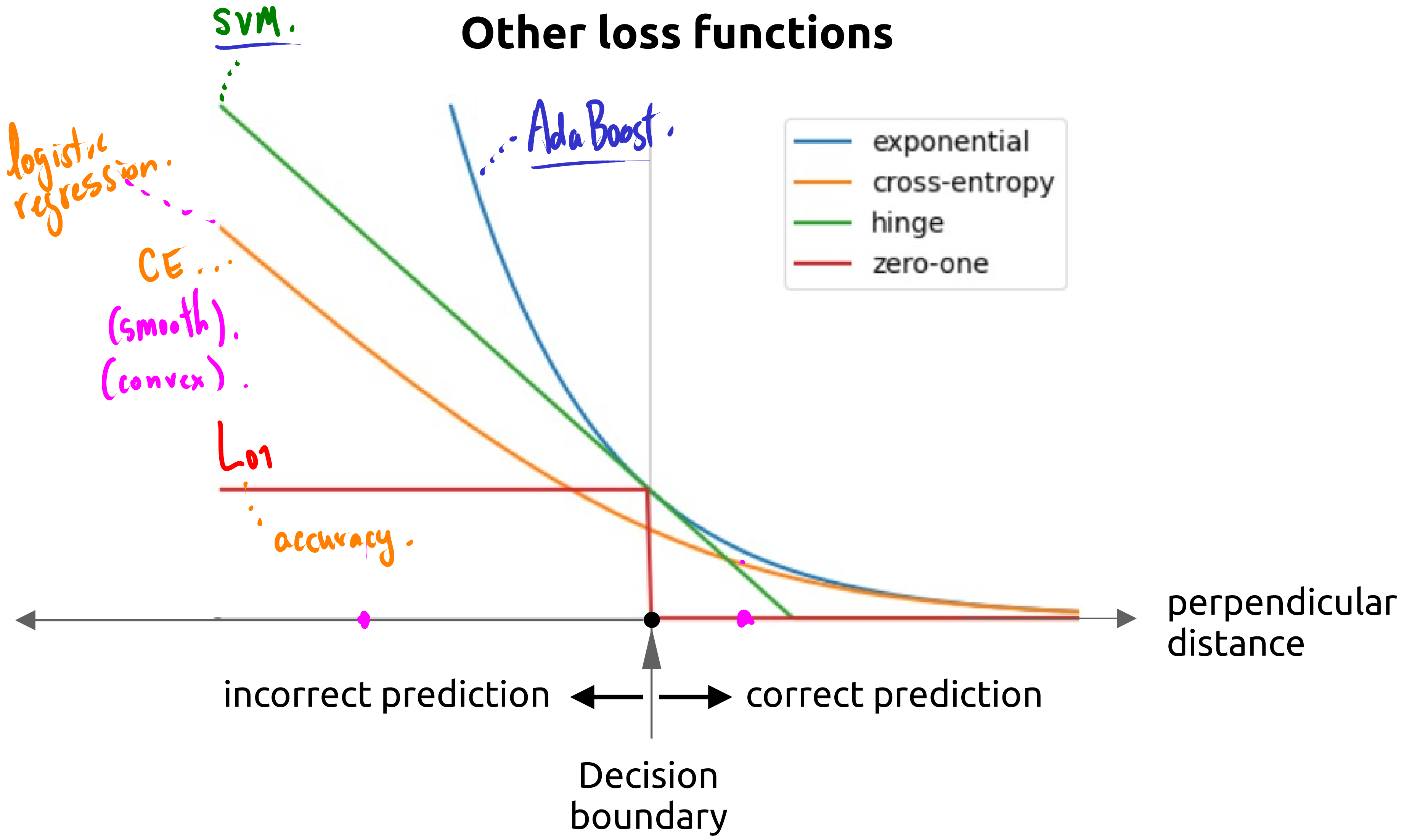
loss for $\hat{y}=1$

$z = \theta_0 + x^T \theta_1$

Decision boundary

$x \cdots y=1$

Other loss functions



Multi-class logistic regression

2-class logistic regression boundary for 1 vs K

- $\Omega_Y = \{c_1, \dots, c_K\}$
- c_K is a "reference class".
- Solve $K-1$ logstic regressions: $\{c_1, c_K\}, \{c_2, c_K\}, \dots, \{c_{K-1}, c_K\}$.
- Obtain parameter **vectors** $\underline{\hat{\theta}}^{(1)}, \dots, \underline{\hat{\theta}}^{(K-1)}$ (for each logistic regression)

$(\hat{\theta}_0, \hat{\theta}_1) = \underline{\hat{\theta}}^{(1)}$ $\quad \underline{\hat{\theta}}^{(2)}$ $\quad \underline{\hat{\theta}}^{(K-1)}$

2 vs. K

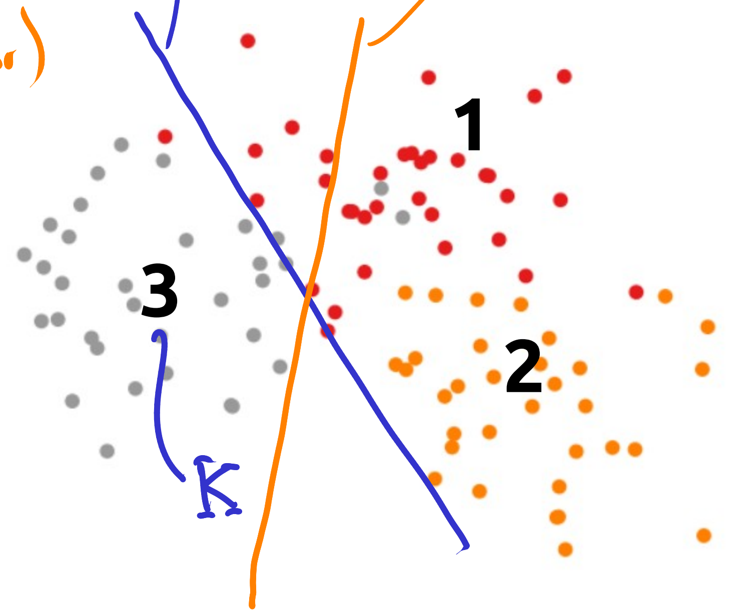
"3"

(added a "1" to x vector)

e.g. $K = 3$

$$\log \frac{p(Y=1 \mid X=x)}{p(Y=3 \mid X=x)} = x^T \underline{\hat{\theta}}^{(1)}$$

$$\log \frac{p(Y=2 \mid X=x)}{p(Y=3 \mid X=x)} = x^T \underline{\hat{\theta}}^{(2)}$$



In general: $$\log \frac{p(Y=c_k \mid X=x)}{p(Y=c_K \mid X=x)} = x^T \underline{\hat{\theta}}^{(k)} \qquad k = 1 \dots K-1$$

exp

$$\therefore \quad p(Y=c_k \mid X=x) = p(Y=c_K \mid X=x)\,\mathsf{exp}(x^T \underline{\hat{\theta}}^{(k)}) \qquad k = 1 \dots K-1 \qquad \texttt{(I)}$$

$$\sum_{k=1}^{K} p(Y=c_k \mid X=x) = 1$$

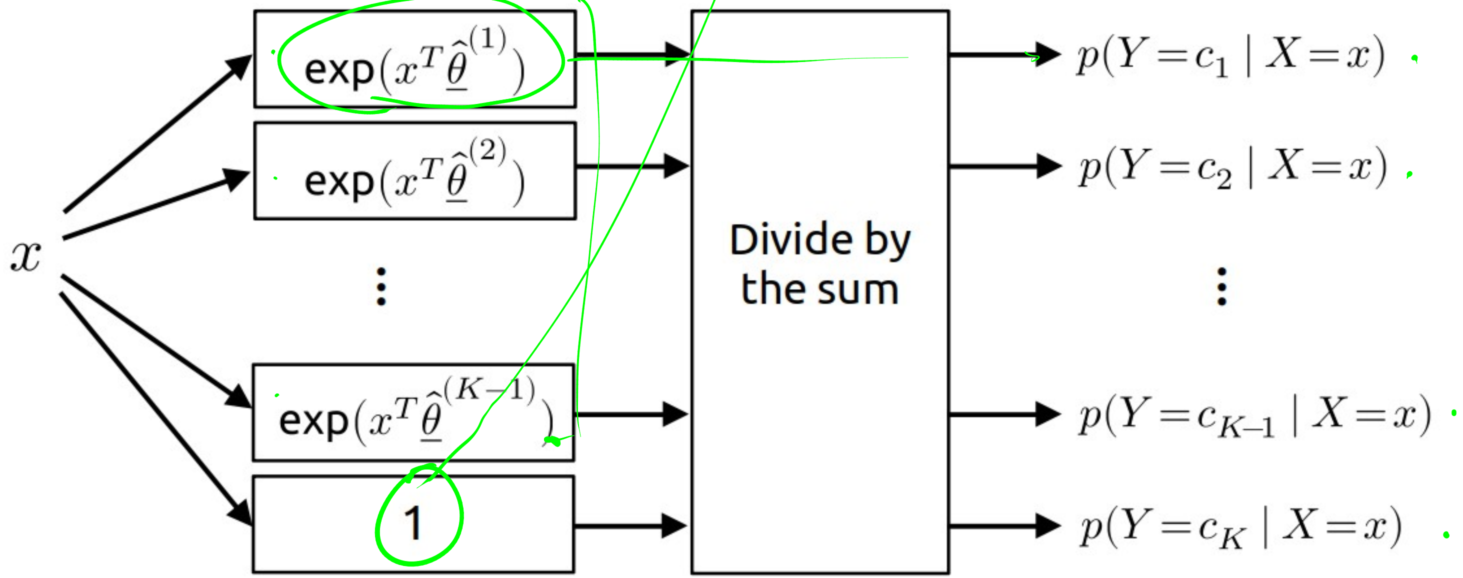
Also: $$p(Y=c_K \mid X=x) = 1 - \sum_{\kappa=1}^{K-1} p(Y=c_\kappa \mid X=x)$$

$$= 1 - p(Y=c_K \mid X=x) \sum_{\kappa=1}^{K-1} \mathsf{exp}(x^T \underline{\hat{\theta}}^{(\kappa)})$$

$$\therefore \quad \boxed{p(Y=c_K \mid X=x) = \frac{1}{1 + \sum_{\kappa=1}^{K-1} \mathsf{exp}(x^T \underline{\hat{\theta}}^{(\kappa)})}} \qquad \texttt{(II)}$$

Softmax

$$p(Y=c_k \mid X=x) = \frac{\exp\left(x^T\hat{\theta}^{(k)}\right)}{1+\sum_{\kappa=1}^{K-1}\exp\left(x^T\hat{\theta}^{(\kappa)}\right)}$$



0.1

0.3

0.05

0.02

1