
From Linear Scores to Class Probabilities

David I. Inouye

| Classification changes **what the score must predict**

Regression: a numerical price, $\hat{y} = w^T x + b$.

Classification: a positive/negative review, with an estimated class probability.

Return to our real Cornell case: **18,746 word features**, 1,200 training reviews.

| Logistic regression follows **model** → **objective** → **solution**

1. **Model — now:** turn the score into a probability.
2. **Objective — next:** measure how well those probabilities predict the labels.
3. **Solution — afterward:** learn the coefficients by numerical optimization.

| A learned score combines **signed contributions**

For one review, let $\boldsymbol{x} \in \mathbb{R}^d$ contain its word features.

$$z = \boldsymbol{w}^\top \boldsymbol{x} + b = \sum_{j=1}^d w_j x_j + b.$$

- x_j : the review's measured feature; **given at prediction time**.
- w_j : how that feature contributes; **learned from labeled reviews**.
- b : a learned offset; z : the resulting scalar score.

With $d = 18,746$, the model learns **18,747 parameters**, not one parameter per review.

| Each contribution is a **weight times a measurement**

Illustrative two-feature review, with **invented weights**:

Feature	Measurement x_j	Weight w_j	Contribution $w_j x_j$
“engaging”	2	+0.8	+1.6
“dull”	1	−1.1	−1.1
Offset	—	$b = -0.2$	−0.2

$$z = 1.6 - 1.1 - 0.2 = 0.3 \implies \hat{y} = 1.$$

The model combines evidence; a word’s contribution depends on **both its value and its weight**.

| A zero threshold divides space with a **hyperplane**

$$\hat{y} = \mathbf{1}\{w^\top x + b \geq 0\}, \quad \text{boundary: } w^\top x + b = 0.$$

- In two dimensions, the boundary is a **line**; in three, a **plane**.
- w determines its orientation; b shifts its position.
- Moving in a direction orthogonal to w leaves the score unchanged.

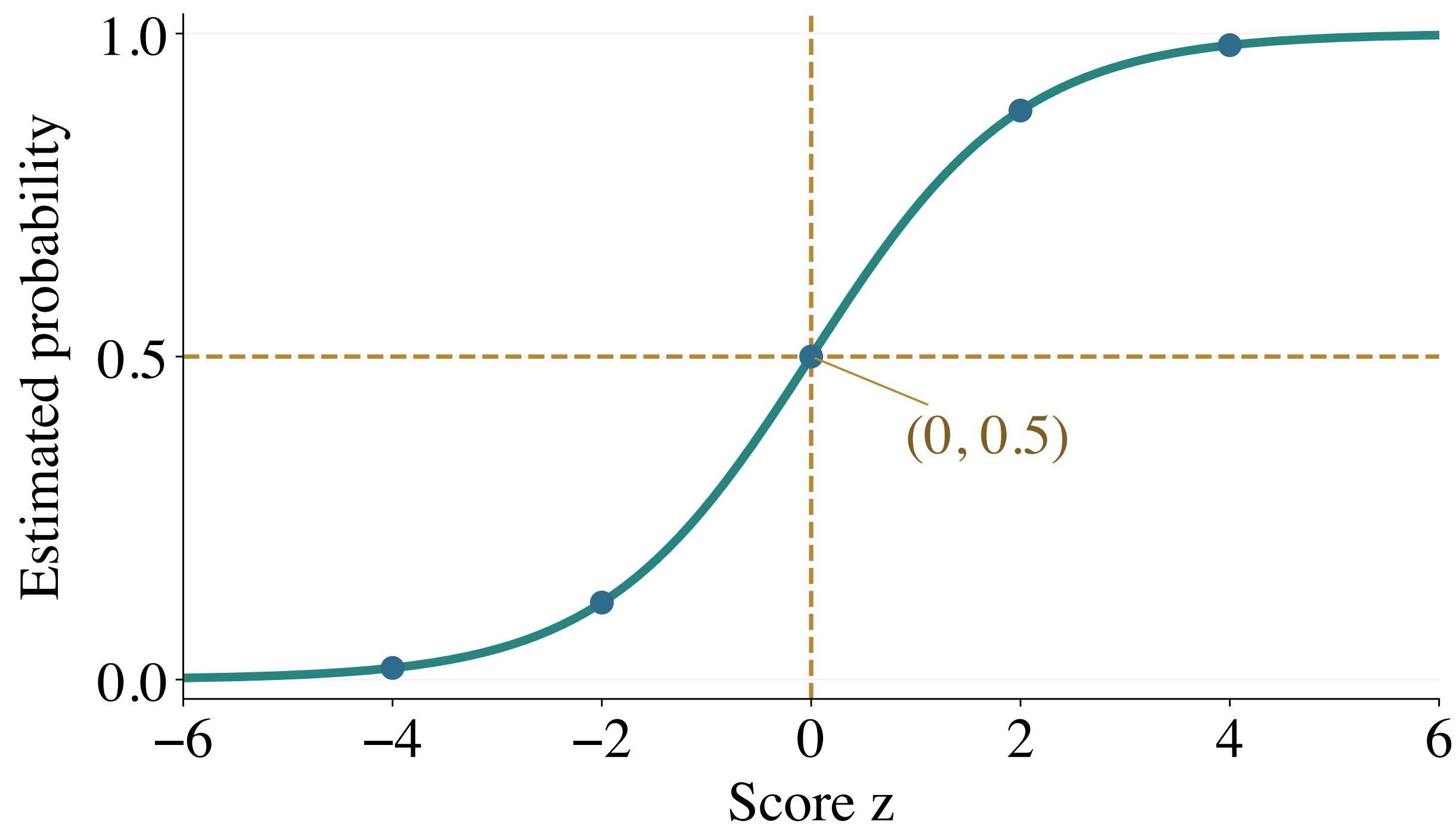
Learning chooses **which direction matters for the labels**, rather than which examples are nearby in every coordinate.

| A sigmoid converts the score into a **probability estimate**

$$\hat{\eta}(x) = \sigma(z) = \frac{1}{1 + e^{-z}}, \quad z = w^T x + b.$$

Score z	Probability of label 1
-4	1.8%
-2	11.9%
0	50%
2	88.1%
4	98.2%

Logistic regression uses this probability model.



| The probability changes smoothly; the **class boundary stays linear**

$$\hat{\eta}(x) \geq \frac{1}{2} \iff w^T x + b \geq 0.$$

- A larger positive score means a larger estimated probability of label 1.
- Multiplying all coefficients by 10 keeps the decisions but makes probabilities more extreme.
- An S-shaped probability function does **not** make this decision boundary curved.

| The probability model is fixed; now choose the fitting objective

Model: a linear score followed by a sigmoid.

Now: what loss should tell us whether its probabilities fit the labels?

Next: solve for the coefficients that minimize that loss.

| Accuracy gives little guidance **between boundary crossings**

For one positive review, compare scores $z = -3$ and $z = -0.1$.

Score	Predicted probability	Predicted class	Classification error
-3	4.7%	0	1
-0.1	47.5%	0	1

The probability moved toward the observed label, but accuracy sees **no improvement yet**.

A smooth loss can provide information about such changes before the class flips.

| Log loss rewards probability assigned to the **observed outcome**

Let $a = \hat{\eta}(x)$ and $y \in \{0, 1\}$.

$$\ell(y, a) = -y \log a - (1 - y) \log(1 - a) = \begin{cases} -\log a, & y = 1, \\ -\log(1 - a), & y = 0. \end{cases}$$

Give the observed outcome high probability $\rightarrow$ **small loss**.

Give the observed outcome nearly zero probability $\rightarrow$ **large loss**.

| Confident mistakes receive a larger penalty

For a positive outcome, $y = 1$:

Predicted probability a	Class at threshold 0.5	Log loss — $\log a$
0.99	Correct	0.010
0.60	Correct	0.511
0.40	Wrong	0.916
0.01	Wrong	4.605

Unlike classification error, log loss distinguishes **how confidently** a prediction was made.

| Would accuracy and log loss choose the same model?

Two observed outcomes are both positive: $y_1 = y_2 = 1$.

Model	Predicted probabilities	Accuracy	Average log loss
A	(0.60, 0.60)	100%	0.511
B	(0.99, 0.49)	50%	0.362

What does this comparison establish?

- A. Accuracy and log loss always rank models identically.
- B. Better log loss can coexist with worse thresholded accuracy.
- C. Model B must generalize better to every new population.

B. The criteria assess different outputs. Neither two-example score establishes population performance.

| MSE and log loss express **different prediction criteria**

Numerical regression	Logistic probability prediction
Output: $\hat{y} = z$	Output: $a = \sigma(z)$
Loss: $\frac{1}{2}(z - y)^2$	Loss: $-y \log a - (1 - y) \log(1 - a)$
Penalize numerical residual size.	Penalize low probability of the observed label.

Squared error on probabilities is also possible, but composing it with a sigmoid generally loses convexity **in the coefficients**.

| The logistic objective has **no general closed-form coefficient solution**

With $z_i = \mathbf{w}^\top \mathbf{x}_i + b$ and $a_i = \sigma(z_i) = e^{z_i} / (1 + e^{z_i})$:

$$\log a_i = z_i - \log(1 + e^{z_i}), \quad \log(1 - a_i) = -\log(1 + e^{z_i}).$$

$$\ell(y_i, a_i) = -y_i \log a_i - (1 - y_i) \log(1 - a_i)$$

(Usual binary log loss)

$$= -y_i z_i + [y_i + (1 - y_i)] \log(1 + e^{z_i})$$

(Substitute the two log identities)

$$= \log(1 + e^{z_i}) - y_i z_i$$

(Combine the coefficients)

Average the same loss over observations:

$$J(\mathbf{w}, b) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, a_i) = \frac{1}{n} \sum_{i=1}^n [\log(1 + e^{z_i}) - y_i z_i].$$

OLS gave linear stationary equations; logistic loss gives **nonlinear equations** in the coefficients.

Next: what does the objective's shape guarantee?