Maximum Likelihood Estimation

and

Generalized Linear Models

Prepared by: Joseph Bakarji

Probabilistic Interpretation of Linear Regression

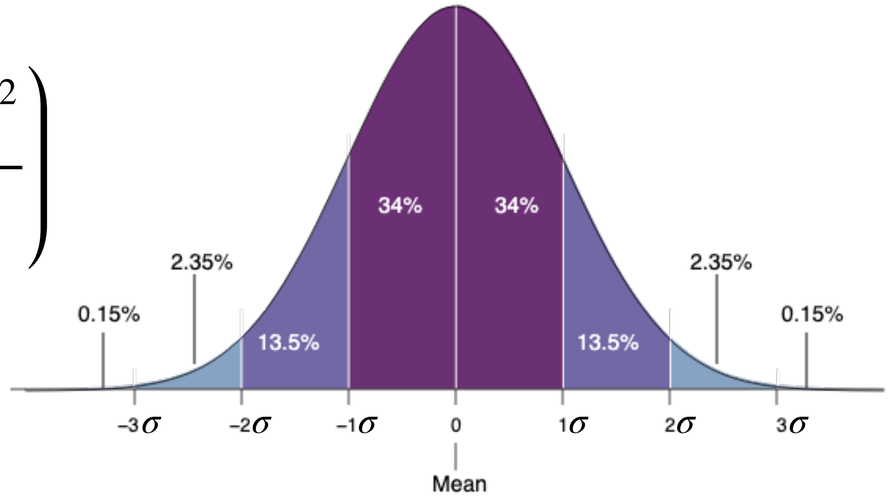
Assume noise is normally distributed around model

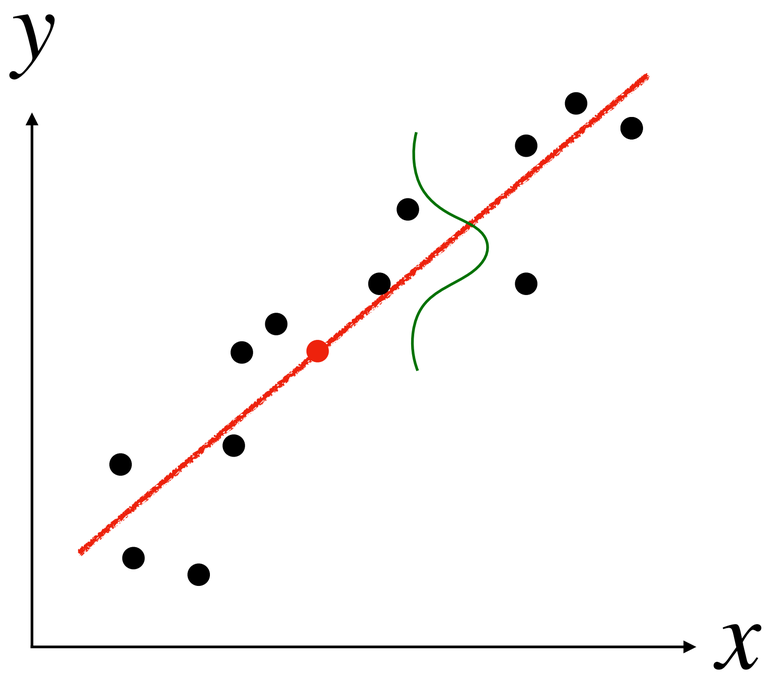
$$y^{(i)} = \theta^\top x^{(i)} + \varepsilon^{(i)}$$

Normally distributed

$$\mathcal{N}\left(0, \sigma^2\right)$$

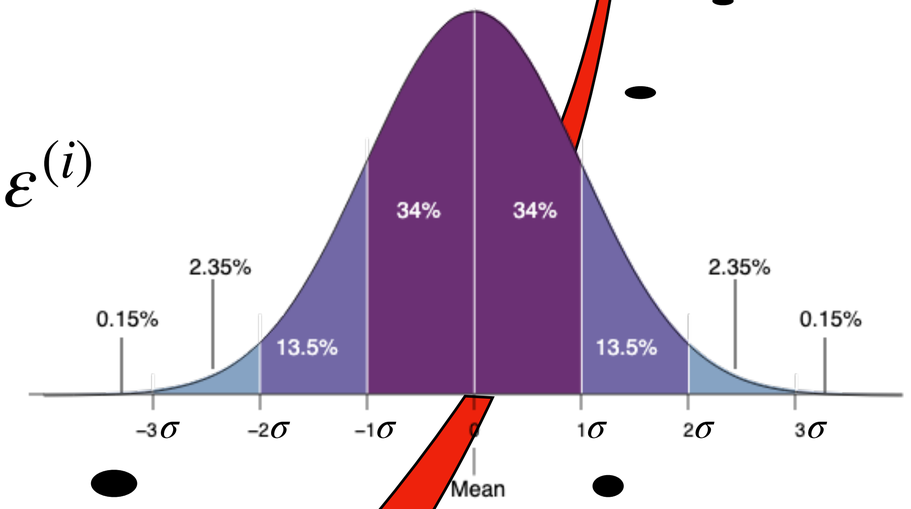
$$p\left(\varepsilon^{(i)}\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(\varepsilon^{(i)}\right)^2}{2\sigma^2}\right)$$

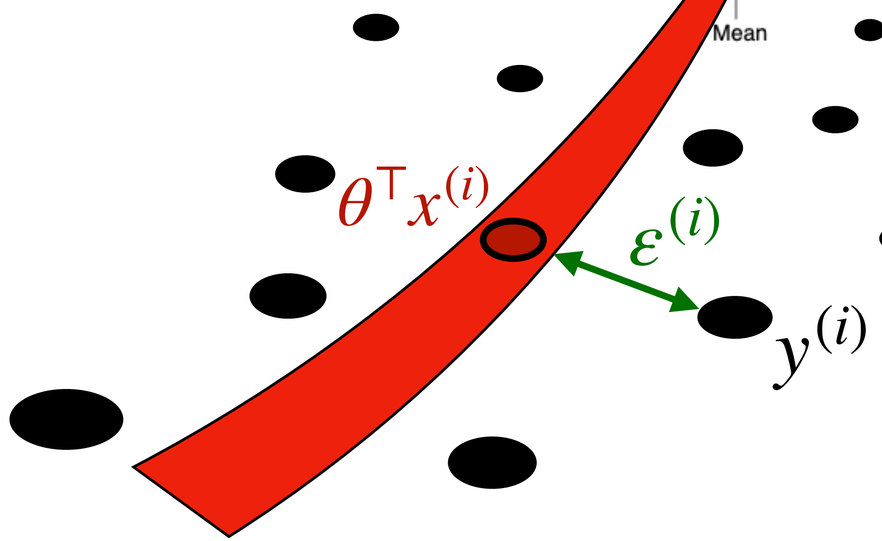


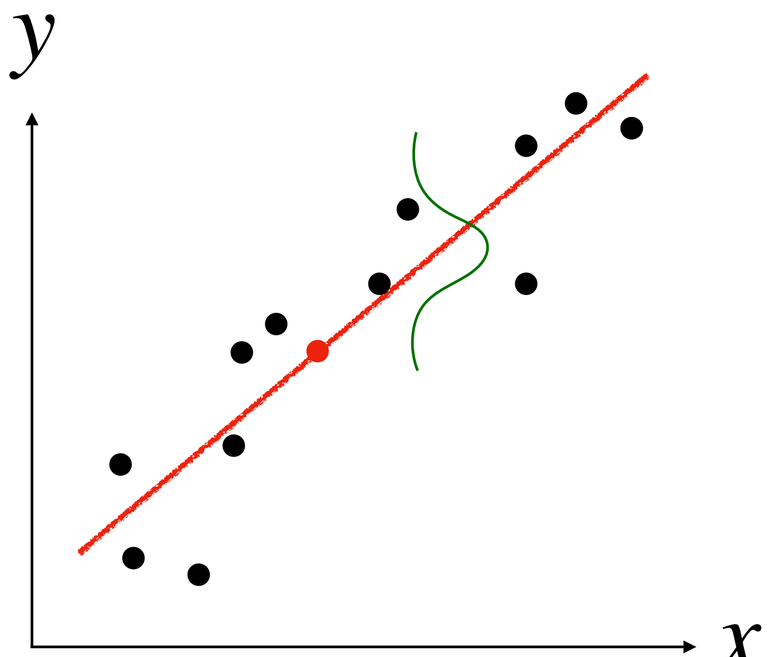


Probabilistic Interpretation of Linear Regression

$$y^{(i)} = \theta^\top x^{(i)} + \varepsilon^{(i)}$$







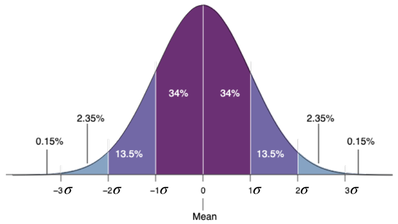
https://www.vecteezy.com/vector-art/19483322-slacklining-vector-illustration-design

Probabilistic Interpretation of Linear Regression

Assume noise is normally distributed around model

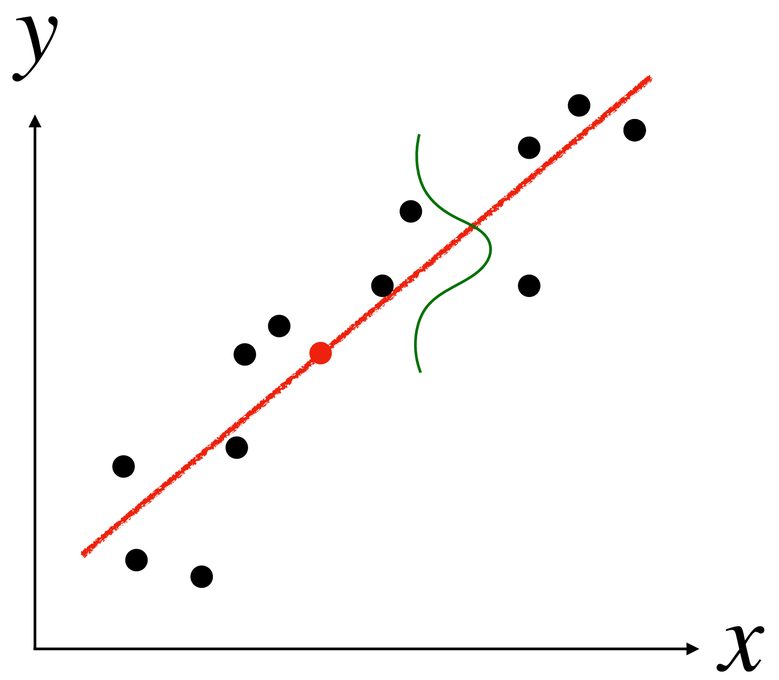
$$y^{(i)} = \theta^\top x^{(i)} + \varepsilon^{(i)}$$

$$\mathcal{N}\left(0, \sigma^2\right)$$



$$p\left(\varepsilon^{(i)}\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(\varepsilon^{(i)}\right)^2}{2\sigma^2}\right)$$

$$p\left(y^{(i)} \mid x^{(i)}; \theta\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y^{(i)} - \theta^\top x^{(i)}\right)^2}{2\sigma^2}\right)$$

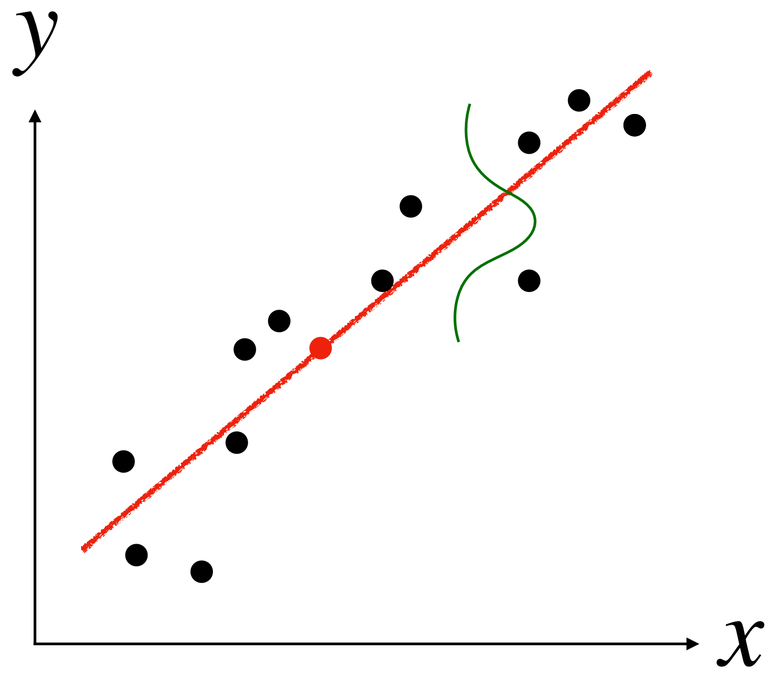


Likelihood of output given input

$$L(\theta) = \prod_{i=1}^{n} p\left(y^{(i)} \,|\, x^{(i)}; \theta\right)$$

Independent and Identically Distributed (IID)

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y^{(i)} - \theta^{\top} x^{(i)}\right)^2}{2\sigma^2}\right)$$



Log-likelihood

$$\mathscr{L}(\theta) = \log L(\theta)$$

$$= \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y^{(i)} - \theta^{\top} x^{(i)}\right)^2}{2\sigma^2}\right)$$

$$= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y^{(i)} - \theta^{\top} x^{(i)}\right)^2}{2\sigma^2}\right) = n \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y^{(i)} - \theta^{\top} x^{(i)}\right)^2$$

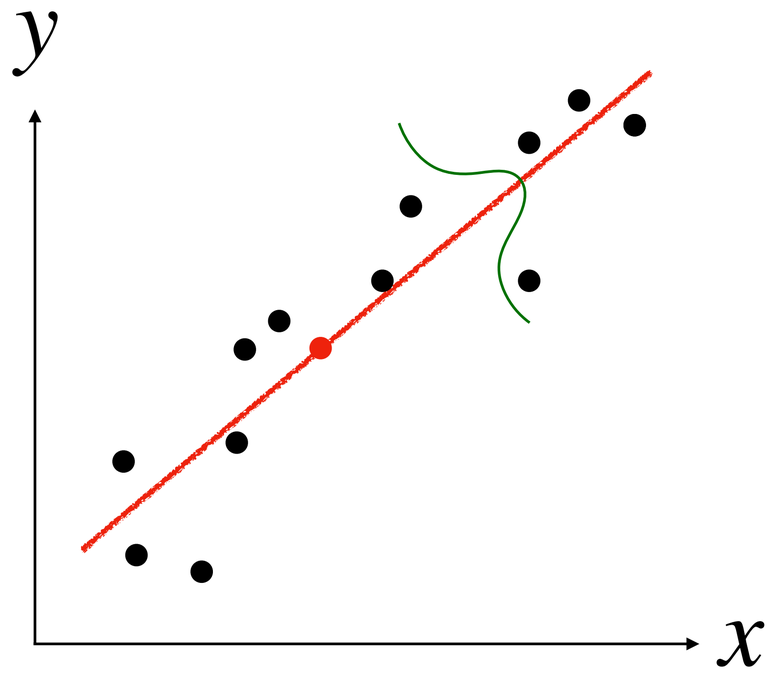
Maximize Log-likelihood

$$\mathscr{L}(\theta) = \log L(\theta)$$

$$= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y^{(i)} - \theta^\top x^{(i)}\right)^2}{2\sigma^2} \right)$$

$$= n \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y^{(i)} - \theta^\top x^{(i)}\right)^2$$

Maximize $\mathscr{L}(\theta)$ ⟶ **Minimize** $\frac{1}{2} \sum_{i=1}^{n} \left(y^{(i)} - \theta^\top x^{(i)}\right)^2$

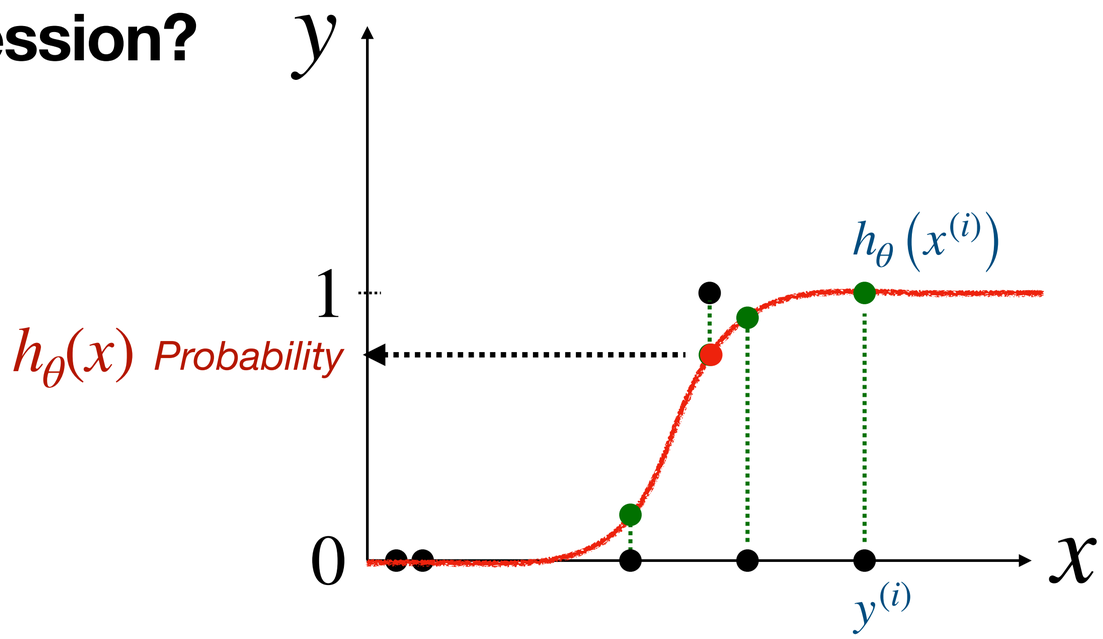


What if the noise is not Gaussian?

Why not Least Squares in Logistic Regression?

$$y = h_\theta(x)$$

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



Probability of output given input

$$P\left(y = 1 \,\middle|\, x; \theta\right) = h_\theta(x)$$

$$P\left(y = 0 \,\middle|\, x; \theta\right) = 1 - h_\theta(x)$$

True label

$$p(y \,|\, x; \theta) = \left(h_\theta(x)\right)^{y} \left(1 - h_\theta(x)\right)^{1-y}$$

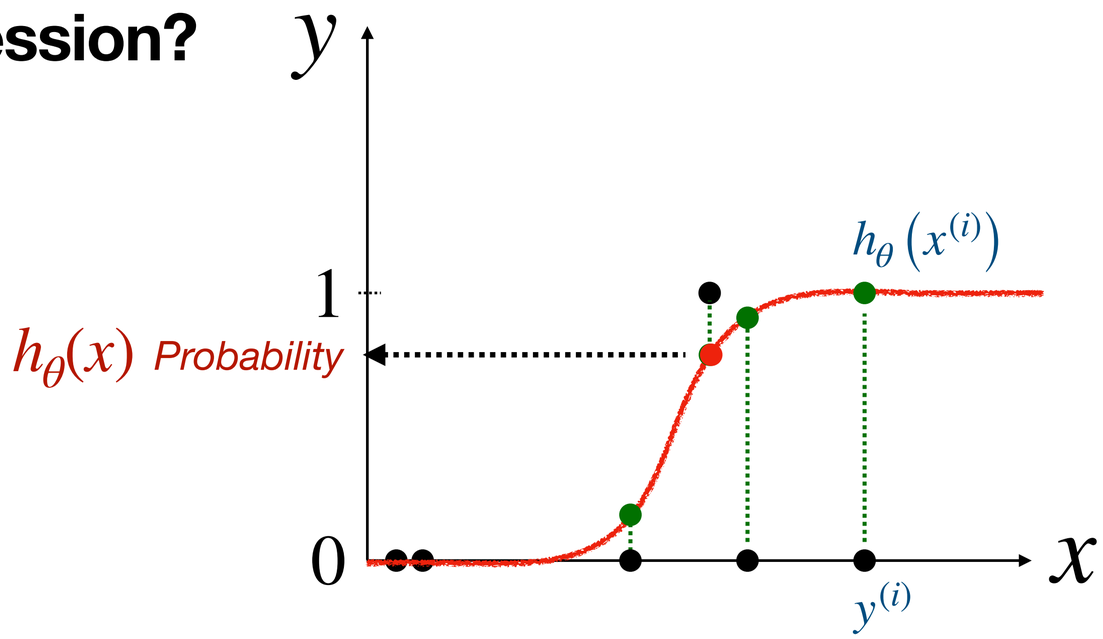
Likelihood!

For Bernoulli Distributed Noise

Why not Least Squares in Logistic Regression?

$$y = h_\theta(x)$$

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



Probability of output given input

$$P\left(y = 1 \,\middle|\, x; \theta\right) = \sigma(\theta^\top x)$$

$$P\left(y = 0 \,\middle|\, x; \theta\right) = 1 - \sigma(\theta^\top x)$$

$$\longrightarrow$$

True label

$$p(y \,|\, x; \theta) = \left(\sigma(\theta^\top x)\right)^{y} \left(1 - \sigma(\theta^\top x)\right)^{1-y}$$

Likelihood!

For Bernoulli Distributed Noise

Bernoulli Distribution

Properties [edit]

If X is a random variable with a Bernoulli distribution, then:

$$\Pr(X = 1) = p = 1 - \Pr(X = 0) = 1 - q.$$

The probability mass function f of this distribution, over possible outcomes *k*, is

$$f(k;p) = \begin{cases} p & \text{if } k = 1, \\ q = 1 - p & \text{if } k = 0. \end{cases}$$ [3]

This can also be expressed as

$$f(k;p) = p^k(1-p)^{1-k} \quad \text{for } k \in \{0, 1\}$$

or as

$$f(k;p) = pk + (1-p)(1-k) \quad \text{for } k \in \{0, 1\}.$$

The Bernoulli distribution is a special case of the binomial distribution with $n = 1$.[4]

https://en.wikipedia.org/wiki/Bernoulli_distribution

Define Log-likelihood

Likelihood

$$p(y \mid x; \theta) = \left(h_\theta(x)\right)^{y} \left(1 - h_\theta(x)\right)^{1-y}$$

for all (x, y) pair

$$L(\theta) = \prod_{i=1}^{n} p\left(y^{(i)} \mid x^{(i)}; \theta\right)$$

$$= \prod_{i=1}^{n} h_\theta\left(x^{(i)}\right)^{y^{(i)}} \left(1 - h_\theta(x^{(i)})\right)^{1-y^{(i)}}$$

log

$$\mathscr{L}(\theta) = \sum_{i=1}^{n} y^{(i)} \log h_\theta\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - h_\theta\left(x^{(i)}\right)\right)$$

Maximize Log-likelihood

$$\mathscr{L}(\theta) = \sum_{i=1}^{n} y^{(i)} \log h_\theta\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - h_\theta\left(x^{(i)}\right)\right)$$

Update rule?

while not converged:

$$\theta_j := \theta_j + \alpha \frac{\partial \mathscr{L}(\theta)}{\partial \theta_j}$$

Derive

Gradient Descent

for t = 1...T:

  for all parameters j:

$$\theta_j := \theta_j - \alpha \sum_{i=1}^{n} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x_j^{(i)}$$

Maximize Log-likelihood

$$\mathscr{L}(\theta) = \sum_{i=1}^{n} y^{(i)} \log h_\theta\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - h_\theta\left(x^{(i)}\right)\right)$$

Update rule

while not converged:

$$\theta := \theta + \alpha \nabla_\theta \mathscr{L}(\theta)$$

Derive

Gradient Descent

for t = 1...T:

$$\theta := \theta - \alpha \sum_{i=1}^{n} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x^{(i)}$$

🤯 **Same as linear regression** 🤔

Generalized Linear Models

Gaussian Distribution → Linear Regression

Bernoulli Distribution → Logistic Regression

Update rule

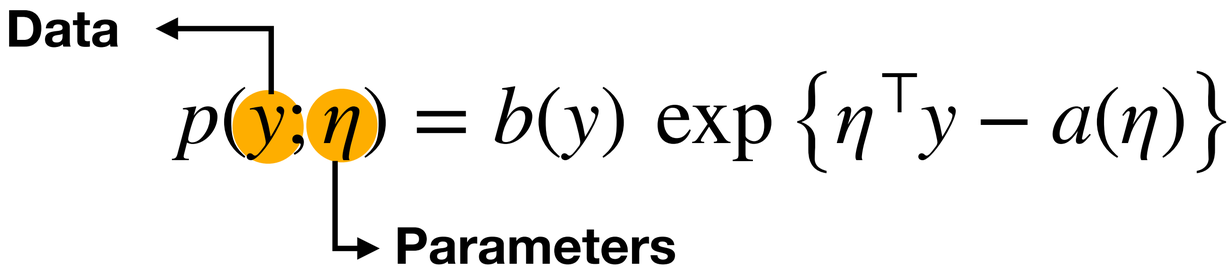
$$\theta := \theta - \alpha \sum_{i=1}^{n} \left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right) x^{(i)}$$

Are there other distributions that lead to the same update rule?

Exponential Family

Family of distributions for which we can derive the same update rule

Assumption: $p(y \mid x; \theta)$ is an exponential family



- $b(y)$ is called the base measure (not depend on η)
- $a(\eta)$ is called the log partition function (not depend on y)
- $a(\eta)$, y and $b(y)$ are scalar. η and y have the same dimensions.

Example 1: Bernoulli Distribution -> Logistic Regression

Data

$$p(y;\eta) = b(y)\ \exp\left\{\eta^\top y - a(\eta)\right\}$$

Natural Parameters

Bernoulli Distribution

$$p(y;\phi) = \phi^y(1-\phi)^{1-y} = \exp\left\{ y \log\frac{\phi}{1-\phi} + \log(1-\phi) \right\}$$

η $\quad$ $a(\eta)$

Show that term is only a function of η

Example 2: Gaussian Distribution -> Linear Regression

Data

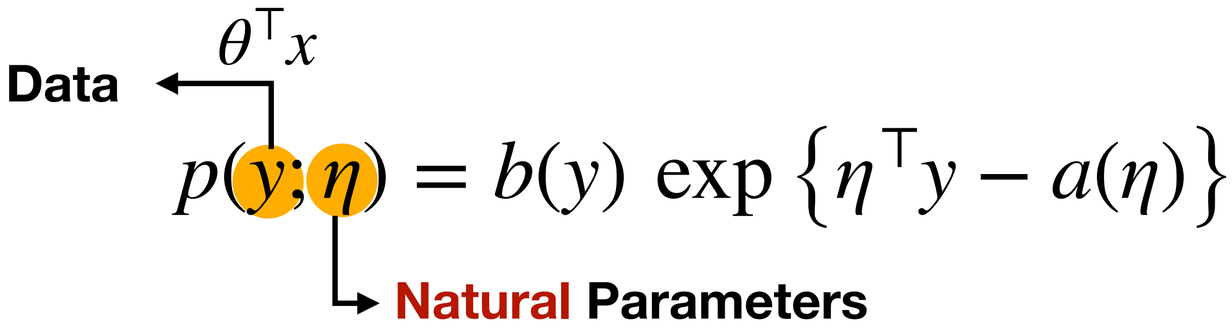
$$p(y;\eta) = b(y)\ \exp\left\{\eta^\top y - a(\eta)\right\}$$

Natural Parameters

Gaussian Distribution

$$p(y;\mu) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(y-\mu)^2\right\} = \underbrace{\frac{1}{\sqrt{2\pi}} e^{-y^2/2}}_{b(y)} \left\{ \underbrace{\mu}_{\eta} y - \underbrace{\frac{1}{2}\mu^2}_{a(\eta)} \right\}$$

Why do we care?



Inference is Easy: $E[y;\eta] = \dfrac{da(\eta)}{d\eta}$ $Var[y;\eta] = \dfrac{d^2a(\eta)}{d\eta^2}$

Learning is Easy: **Maximum Likelihood Estimation** leads to **convex** problem in η

Generalized Linear Models

Assumption: $p(y \mid x; \theta)$ is an exponential family

Data Type → Probability Distribution

Binary → Bernoulli ⟶ **Logistic Regression**

Real → Gaussian ⟶ **Linear Regression**

Counts → Poisson

Positive Real → Gamma, Exponential

Distributions → Dirichlet

Generalized Linear Models

Assumption: $p(y \mid x; \theta)$ is an exponential family

The natural parameter is linear in the inputs

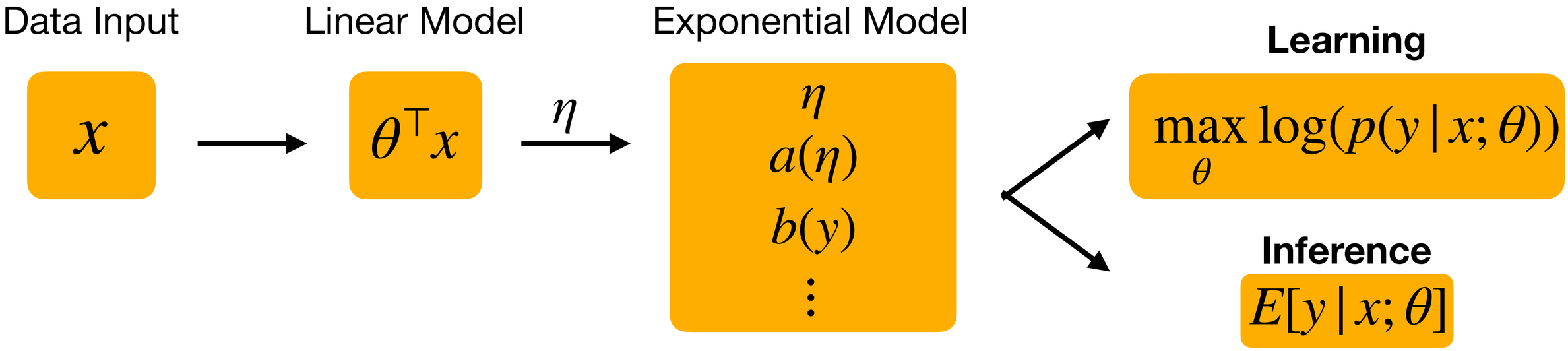
$$\eta = \theta^\top x$$

Predictor is a natural consequence

$$h_\theta(x) = E[y \mid x; \theta]$$

Generalized Linear Models

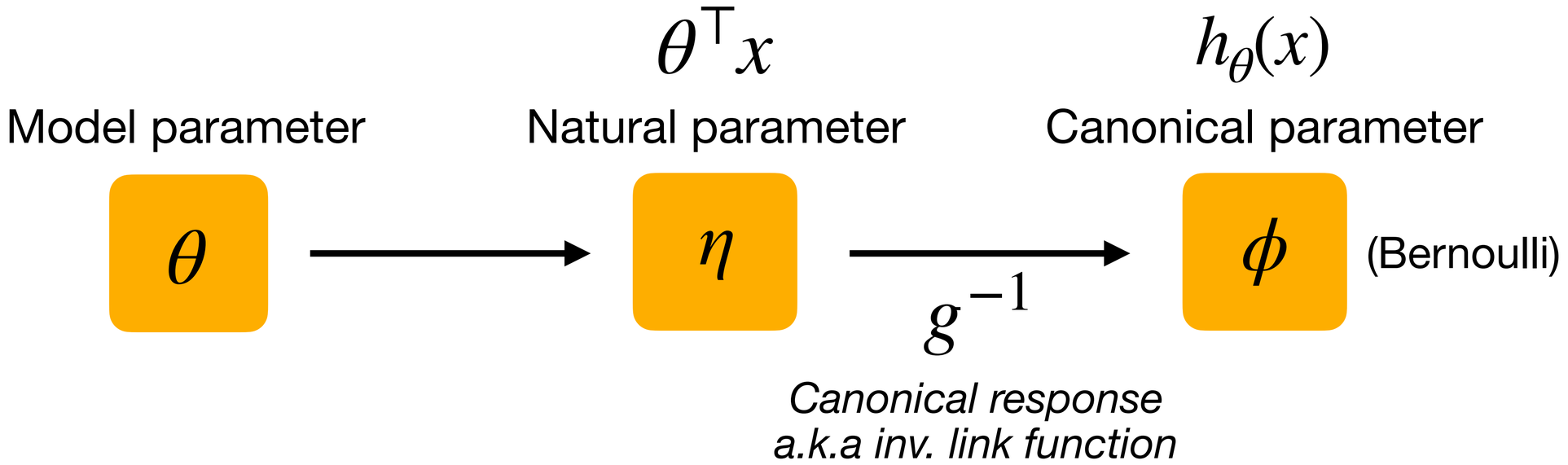
Assumption: $p(y \mid x; \theta)$ is an exponential family



Update Rule:

$$\theta := \theta - \alpha \sum_{i=1}^{n} \left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right) x^{(i)}$$

Terminology



Logistic Regression: $h_\theta(x) = E[y \mid x; \theta]$

$$\phi = \frac{1}{1 + e^{-\eta}} = \frac{1}{1 + e^{-\theta^\top x}}$$

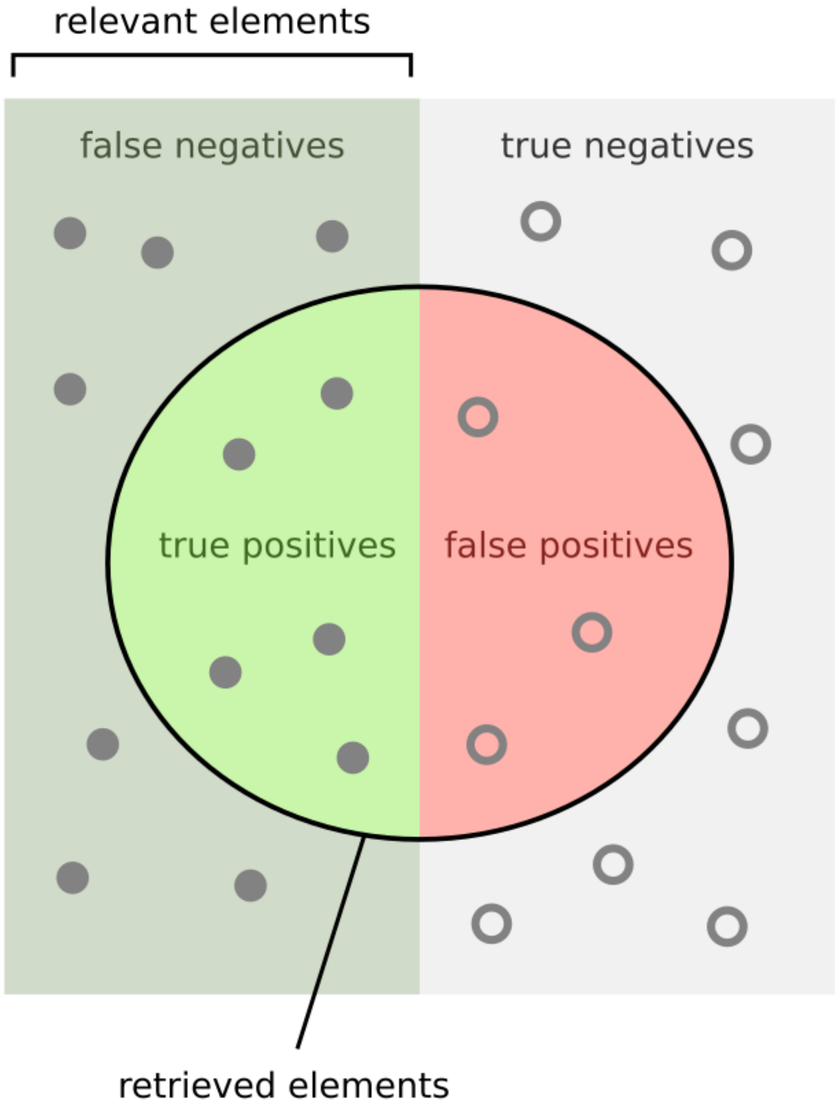
Evaluation Metrics

Sources: [4][5][6][7][8][9][10][11] view · talk · edit

		Predicted condition			
	Total population $= P + N$	**Predicted positive**	**Predicted negative**	Informedness, bookmaker informedness (BM) $= TPR + TNR - 1$	Prevalence threshold (PT) $= \frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	**Positive (P)** [a]	**True positive** (TP), hit[b]	**False negative** (FN), miss, underestimation	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate type II error [c] $= \frac{FN}{P} = 1 - TPR$
	Negative (N)[d]	**False positive** (FP), false alarm, overestimation	**True negative** (TN), correct rejection[e]	False positive rate (FPR), probability of false alarm, fall-out type I error [f] $= \frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$
	Prevalence $= \frac{P}{P + N}$	Positive predictive value (PPV), precision $= \frac{TP}{TP + FP} = 1 - FDR$	Negative predictive value (NPV) $= \frac{TN}{TN + FN} = 1 - FOR$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$
	Accuracy (ACC) $= \frac{TP + TN}{P + N}$	False discovery rate (FDR) $= \frac{FP}{TP + FP} = 1 - PPV$	False omission rate (FOR) $= \frac{FN}{TN + FN} = 1 - NPV$	Markedness (MK), deltaP (Δp) $= PPV + NPV - 1$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$
	Balanced accuracy (BA) $= \frac{TPR + TNR}{2}$	F_1 score $= \frac{2\,PPV \times TPR}{PPV + TPR} = \frac{2\,TP}{2\,TP + FP + FN}$	Fowlkes–Mallows index (FM) $= \sqrt{PPV \times TPR}$	*phi* or Matthews correlation coefficient (MCC) $= \sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$	Threat score (TS), critical success index (CSI), Jaccard index $= \frac{TP}{TP + FN + FP}$

https://en.wikipedia.org/wiki/Precision_and_recall

Precision and Recall



How many retrieved items are relevant?

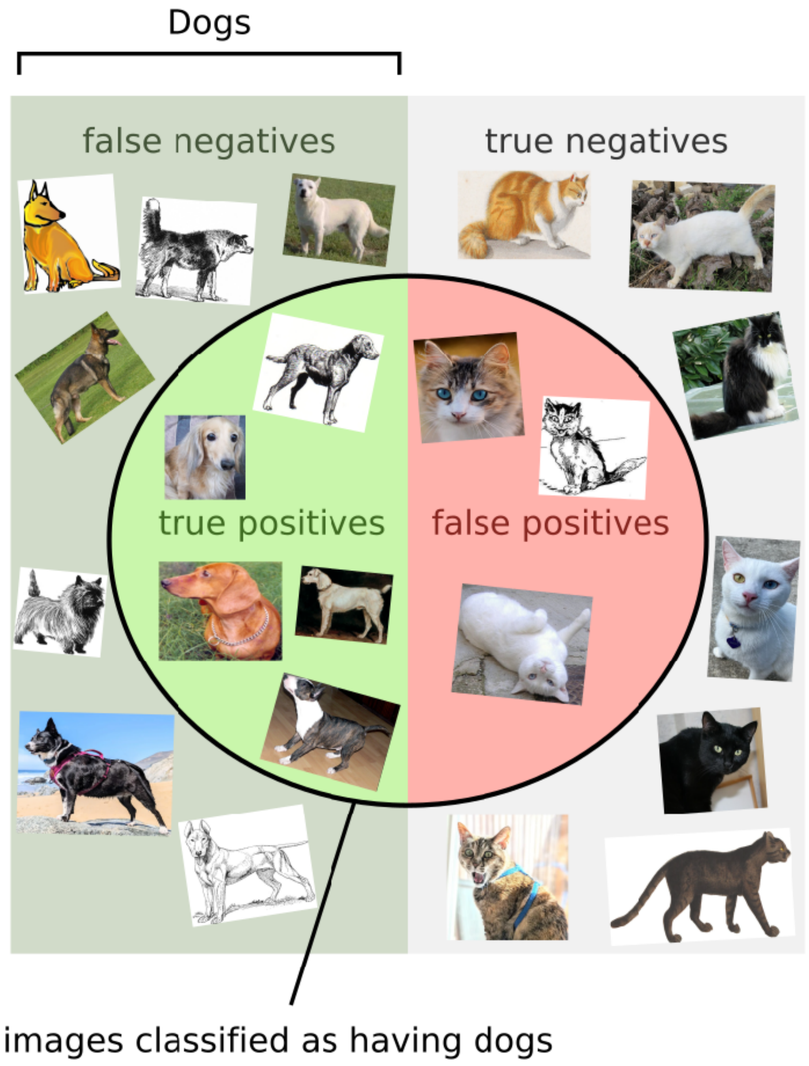
$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are retrieved?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

https://en.wikipedia.org/wiki/Precision_and_recall

Precision and Recall



$$\text{Precision} = \frac{\text{5 true pos.}}{\text{8 total pos.}} \qquad \text{Recall} = \frac{\text{5 true pos.}}{\text{12 total dogs}}$$

$$\text{Prevalence} = \frac{\text{12 total dogs}}{\text{22 total images}}$$

$$\text{Accuracy} = \frac{\text{5 true pos. + 7 true neg.}}{\text{22 total images}}$$

$$F_1 = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

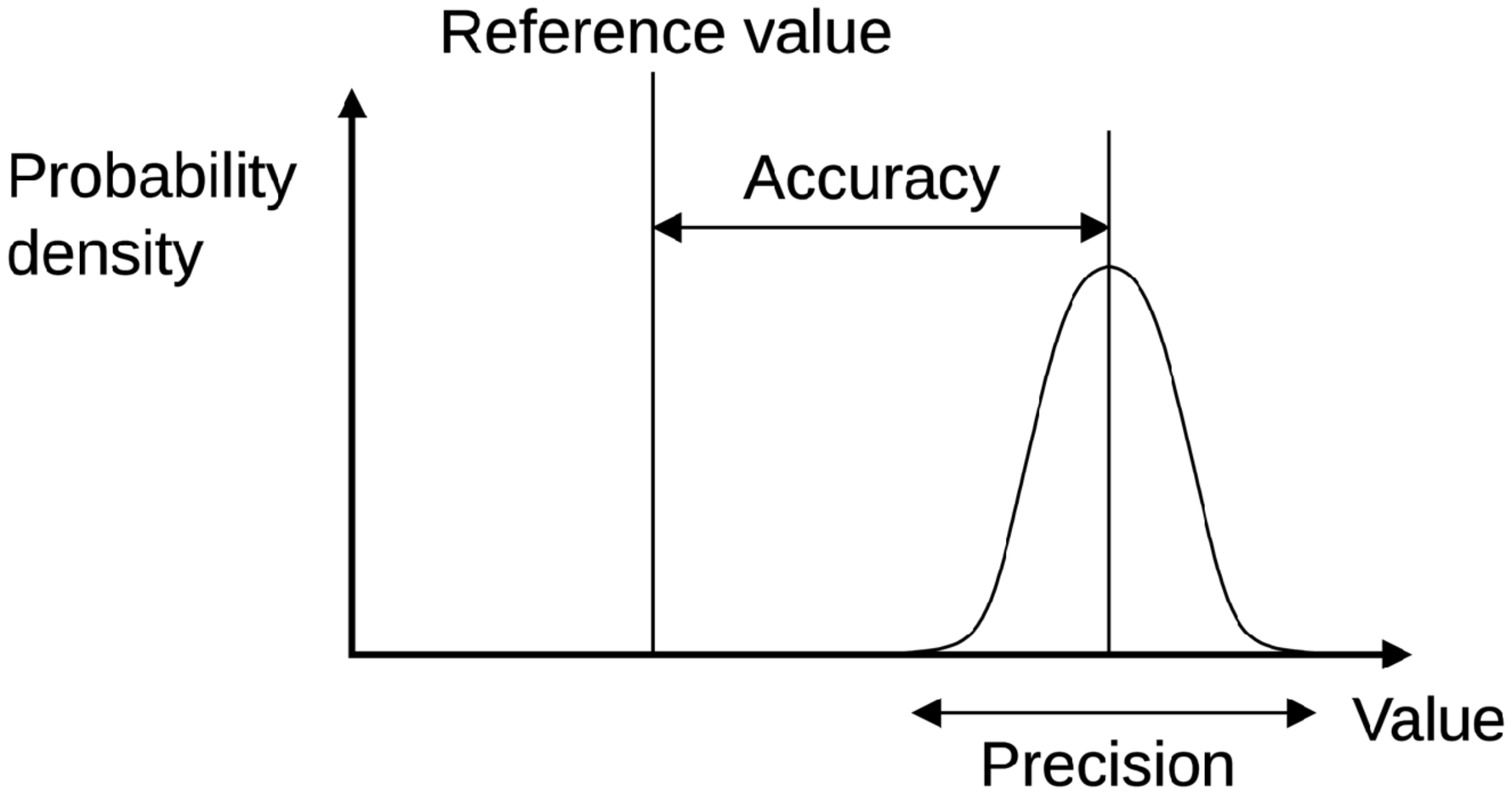
```
sklearn.metrics.f1_score(y_true, y_pred, *, labels=None, pos_label=1,
average='binary', sample_weight=None, zero_division='warn') #          [source]
```

Compute the F1 score, also known as balanced F-score or F-measure.

The F1 score can be interpreted as a harmonic mean of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. The formula for the F1 score is:

https://en.wikipedia.org/wiki/Precision_and_recall

Accuracy and Precision



https://en.wikipedia.org/wiki/Precision_and_recall