

Statistics and machine learning

Probability

$$P(A | B) = \frac{P(B | A)P(A)}{P(B)} \quad P(B) = \sum_i P(B | A_i)P(A_i) \quad \mathbb{E}[X] = \sum_x x P(X = x) = \int_{-\infty}^{\infty} x f(x) dx$$

The normal distribution and its standardization:

$$f(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) \quad Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

$$\text{Var}(X) = \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \quad \text{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$$

Estimation

Sample mean, sample variance and a 95 % confidence interval for the mean:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad \bar{x} \pm t_{n-1, 0.975} \frac{s}{\sqrt{n}}$$

Maximum likelihood and ordinary least squares:

$$\hat{\theta}_{\text{ML}} = \arg \max_{\theta} \prod_{i=1}^n f(x_i | \theta) = \arg \max_{\theta} \sum_{i=1}^n \log f(x_i | \theta) \quad \hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

The covariance matrix of three variables, with the correlation coefficient:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 \end{pmatrix} \quad \rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Information

$$H(X) = -\sum_i p_i \log_2 p_i \quad D_{\text{KL}}(P \parallel Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)} \quad I(X; Y) = H(X) - H(X | Y)$$

Machine learning

$$\text{softmax}(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}} \quad \sigma(x) = \frac{1}{1 + e^{-x}} \quad \text{ReLU}(x) = \max(0, x)$$

Cross-entropy loss over N examples and the gradient step with learning rate η :

$$\mathcal{L}(\theta) = -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K y_{ik} \log \hat{y}_{ik} \quad \theta_{t+1} = \theta_t - \eta \nabla_{\theta} \mathcal{L}(\theta_t)$$

Adam, with bias-corrected moments:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \quad \hat{m}_t = \frac{m_t}{1 - \beta_1^t} \quad \theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \quad \hat{v}_t = \frac{v_t}{1 - \beta_2^t}$$

Scaled dot-product attention and a layer of a feed-forward network:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \quad \mathbf{h}^{(l+1)} = \phi(W^{(l)}\mathbf{h}^{(l)} + \mathbf{b}^{(l)})$$

A confusion matrix and the scores read from it:

	Predicted +	Predicted −			
Actual +	TP	FN	Precision = $\frac{TP}{TP + FP}$	Recall = $\frac{TP}{TP + FN}$	$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
Actual −	FP	TN			