Probability Formulas

Discrete Random Variables

Properties of Probability Mass Function (PMF)	$0 \leq p_X(x) \leq 1$ $\sum_{x\in\mathcal{X}} p_X(x) = 1$ $\sum_{x\in A} p_X(x) = Pr(X \in A)$
Expectation of X	$\mathbb{E}[X] = \sum_{x\in\mathcal{X}} x.p_X(x)$
Properties of Expectation	$\mathbb{E}[a] = a$ $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$ $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
Expectation of a function of a random variable $g(X)$	$\mathbb{E}[g(X)] = \sum_{x\in\mathcal{X}} g(x)p_X(x)$
Mean	$\mu_X = E[X]$
Variance	$Var[X] = \mathbb{E}[(X - \mathbb{E}[X])^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
Variance	$Var(X) = \sum(x_i - \mu)^2 p_X(x_i) = E[(X - \mu)^2]$
Variance (better)	$Var(X) = \sum x_i^2 p_X(x_i) - \mu^2$
Properties of Variance	$Var(kX) = k^2 Var(X)$ $Var(X + b) = Var(X)$
Standard Deviation (SD or SE)	$\mathrm{SD}(X) = \sigma_X = \sqrt{Var(X)}$
Standardized Random Variable (Z)	$Z = \frac{X-\mu}{\sigma}$
Properties of Z	$\mu_Z = 0 \quad \sigma_Z = 1$
Joint distribution (Joint PMF) of X, Y	$p_{XY}(x, y) = Pr(X = x, Y = y)$
Properties of Joint PMF	$\sum_i \sum_j p_{XY}(x_i, y_j) = 1$
Marginal distribution (marginal PMF)	$p_X(x_i) = \sum_j p_{XY}(x_i, y_j)$
Conditional distribution (Conditional PMF)	$p_{Y\|X}(y\|x) = \frac{p_{XY}(x,y)}{p_X(x)}$
Bayes Rule	$P(X\|Y) = P(Y\|X) \cdot \frac{P(X)}{P(Y)}$ $p_{X\|Y}(x\|y) = p_{Y\|X}(y\|x) \cdot \frac{p_X(x)}{p_Y(y)}$
Independence and Joint PMF	$p_{XY}(x, y) = p_X(x) \cdot p_Y(y)$ $\text{ for all } x \in \mathcal{X} \text{ and } y \in \mathcal{Y}$
Independence and Conditional PMF	$p_{X\|Y}(x\|y) = p_X(x)$
Expectation of Joint PMF	$\mathbb{E}[X, Y] = \sum_i \sum_j (x_i, y_j) p_{XY}(x_i, y_j)$
Covariance	$Cov[X, Y] = \mathbb{E}[X, Y] - \mathbb{E}[X]\mathbb{E}[X]$ $Cov[X, Y] = \sum_{i,j} x_i \cdot y_j \cdot p_{XY}(x_i, y_j) - \mu_X \mu_Y$
Variance and Covariance	$Var[X + Y] = Var[X] + Var[Y] + 2Cov[X, Y]$
Joint Expectation, Variance and Covariance of Independent X, Y	$\mathbb{E}[X, Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$ $Var(X + Y) = Var(X) + Var(Y)$ $Cov(X, Y) = 0$
Discrete Cumulative Distribution Function (CDF)	$F_X(x) = Pr(X \leq x) = \sum_{x_i \leq x} p_X(x_i)$

Continuous Random Variables

Properties of Probability Density Function (PDF)	$f_X(x) \geq 0$ $\int_{-\infty}^{\infty} f_X(x) = 1$
Probability of an event and PDF	$Pr(a \leq x \leq b) = \int_a^b f_X(x)dx$
Expectation of X	$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f_X(x)dx$
Expectation of $g(X)$	$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f_X(x)dx$
Linearity of Expectations	$\mathbb{E}[f(X) + g(X)] = \mathbb{E}[f(X)] + \mathbb{E}[g(X)]$
Variance	$Var(X) == \int_{-\infty}^{\infty} x^2 \cdot f_X(x)dx - \mu^2$
Continuous Cumulative Distribution Function (CDF)	$F_X(x) = Pr(X \leq x) = \int_{-\infty}^{\infty} f_X(x)dx$
Probability of Joint Event	$Pr((x, y) \in A) = \int\int_{(x,y)\in A} f_{XY}(x, y)dxdy$
Marginal PDF	$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)dy$
Conditional PDF	$f_{Y\|X}(y\|x) = \frac{f_{XY}(x,y)}{f_X(x)}$
Bayes Rule	$f_{X\|Y}(x\|y) = f_{Y\|X}(y\|x) \cdot \frac{f_X(x)}{f_Y(y)}$
Independence and joint PDF	$f_{XY}(x, y) = f_X(x) \cdot f_Y(y)$ $\text{for all } x, y \in \mathbb{R}$
Independence and conditional PDF	$f_{X\|Y}(x\|y) = f_X(x)$
Joint Expectation	$\mathbb{E}[g(X, Y)] =$ $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x, y)f_{XY}(x, y)dxdy$
Conditional Expectation	$\mathbb{E}(Y\|X = x) = \int_{-\infty}^{\infty} y \cdot f_{Y\|X}(y\|x)dy$

Distribution	PDF or PMF	Mean	Variance
$Bernoulli(p)$	$\begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0. \end{cases}$	p	$p(1-p)$
$Binomial(n, p)$	$\binom{n}{k} p^k(1-p)^{n-k}$ for $0 \leq k \leq n$	np	npq
$Geometric(p)$	$p(1-p)^{k-1}$ for $k = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$Poisson(\lambda)$	$e^{-\lambda}\lambda^x/x!$ for $k = 1, 2, \ldots$	λ	λ
$Uniform(a, b)$	$\frac{1}{b-a}$ $\forall x \in (a, b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$Gaussian(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
$Exponential(\lambda)$	$\lambda e^{-\lambda x}$ $x \geq 0, \lambda > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Regression Formulas

Residual Sum of Squares	$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	Sum of squared errors. Minimized during regression to estimate the coefficients.
Mean Squared Error	$MSE = \frac{RSS}{n}$	Used as a metric of model quality. Can be computed on training set, validation set, or test set.
MSE as an Expectation	$MSE_{Te} = \mathbb{E}(y_0 - \hat{f}(x_0))^2$	
MSE's components	$MSE_{Te} = Var[\hat{f}(x_0)] + (Bias[\hat{f}(x_0)])^2 + Var[\epsilon]$	$MSE = variance + bias^2 + irreducible\ error$
Residual Standard Error	$RSE = \sqrt{\frac{RSS}{n-p-1}}$	Estimate of $\sigma(\epsilon)$. Has units of Y , so not as useful as R^2 which is unit-less. The smaller the RSE, the better the model fits the data.
Irreducible Error	$RSE^2 = \frac{RSS}{n-p-1}$	Estimate of $\sigma^2 = Var(\epsilon)$
Total sum of squares	$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$	Measures the total variance in the response
R^2 statistic	$R^2 = \frac{\text{TSS}-\text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$	Range: $[0, 1]$. The closer to 1, the better model fits the data. For simple linear regression, $R^2 = r^2$. Measures the proportion of variance in response explained by the data.
Coefficients of simple linear regression	$\hat{\beta}1 = \frac{\sum i=1^n (x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$ $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$	Closed form solution for simple linear regression.
Coefficients of linear regression (simple or multiple)	$\beta = (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T \cdot \mathbf{y}$	Closed form solution (linear algebra)
Variances of coefficients for simple linear regression	$\text{Var}\left[\hat{\beta}_1\right] = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i-\bar{x})^2}$ $\text{Var}\left[\hat{\beta}_0\right] = \sigma^2\left(\frac{1}{N} + \frac{\bar{x}^2}{\sum_{i=1}^{N}(x_i-\bar{x})^2}\right)$	$\text{SD}(\hat{\beta}_1)^2$ $\text{SD}(\hat{\beta}_0)^2$
$t-$ statistic	$t = \frac{\hat{\beta}_1}{\text{SD}(\hat{\beta}_1)}$	$t > 2$ rejects the null hypothesis with 95% confidence.
$f-$statistic (for multiple linear regression)	$F = \frac{(\text{TSS}-\text{RSS})/p}{\text{RSS}/(n-p-1)}$	Indicates if any of the coefficients are significant

Classification Formulas

Conditional Class Probability	$p_k(x_i) = Pr(Y = k \| X = x_i)$	$p_1(x_i) = Pr(Y = 1 \| X = x_i)$ $p_0(x_i) = Pr(Y = 0 \| X = x_i)$
Sigmoid function	$\sigma(z) = \frac{1}{1+e^{-z}} = \frac{e^z}{e^z+1}$	$z = \beta_0 x_0 + \beta_1 x_1 + \cdots + \beta_p x_p$
Binary Classification with logistic regression	$p_1(x) = \sigma(\beta_0 + \beta_1 x)$ $= \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$	$p_0(x) = 1 - \sigma(\beta_0 + \beta_1 x)$ $= \frac{1}{1+e^{\beta_0+\beta_1 x}}$
Logit (log odds, z)	$\log\left(\frac{p_1(x)}{1-p_1(x)}\right) = \beta_0 + \beta_1 x$	
z-statistic	$z_1 = \frac{\hat{\beta}_1}{\mathrm{SD}(\hat{\beta}_1)}$	Used exactly as $t-$value is used for linear regression (for hypothesis testing)
Multi-class (multinomial) logistic regression (Softmax)	$p_k(x) = \Pr(Y = k \mid X = x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p}}{\sum_{l=1}^{K} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}$	

LDA Formulas

Conditional Class Probability (Binary)	$Pr(Y = 1 \| X = \mathbf{x}) = \frac{f_1(\mathbf{x})\cdot\pi_1}{f_1(\mathbf{x})\cdot\pi_1 + f_0(\mathbf{x})\cdot\pi_0}$
Conditional class probability (multi-class)	$Pr(Y = k \| X = \mathbf{x}) = \frac{f_k(\mathbf{x})\cdot\pi_k}{\sum_{\ell=1}^{K} f_\ell(\mathbf{x})\pi_\ell}$
Discriminant Function $\delta_k(x)$ for computing $f_k(x)\pi_k$ as a linear function of x	$\delta_k(x) = f_k(x)\pi_k = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$
Estimating LDA parameters $\hat{\mu}_k, \hat{\pi}_k, \hat{\sigma}^2$ from data	$\hat{\pi}_k = \frac{N_k}{N}$ $\hat{\mu}_k = \frac{1}{N_k}\sum_{i:y_i=k} x_i$ $\hat{\sigma}^2 = \sum_{k=1}^{K} \frac{N_k-1}{N-K}\hat{\sigma}_k^2$
Submitting discriminant function to Bayes classifier	$C(x) = \arg\max_k f_k(x)\pi_k$ $C(x) = \arg\max_k \delta_k(x)$

Classification Errors

	Actual Negative	Actual Positive
Predicted Negative	TN	FN
Predicted Positive	FP	TP
	TNR = TN / (TN +FP)	TPR = TP / (TP + FN)
	TNR = Specificity	TPR = Sensitivity
	FPR = 1 - TNR	FNR = 1 - TPR

	FNR	Sensitivity = TPR = 1 - FNR	FPR	Specificity = TNR = 1 - FPR
$\tau \downarrow$	$\downarrow$	$\uparrow$	$\uparrow$	$\downarrow$
$\tau \uparrow$	$\uparrow$	$\downarrow$	$\downarrow$	$\uparrow$