Probability
Experiment: A repeatable procedure that generates an outcome
Sample space: The set of all possible outcomes
Event: An outcome or set of possible outcomes
Conditions for independence of A and B: $P(A\|B) = P(A)$ $P(B\|A) = P(B)$ $P(A \cap B) = P(A)P(B)$
Bayes' Theorem: $P(A\|B) = P(B\|A)P(A)/P(B)$
Mean
Expectation $E[X] = \sum x \cdot p(x)$
Variance $Var(X) = E[(X-\mu)^2] = E[X^2] - \mu^2$
Standard deviation(s.d). $\sigma = \sqrt{Var(X)}$
$\boldsymbol{E[aX+b]} = a \cdot E[X] + b$
$\boldsymbol{E[X+Y]} = E[X] + E[Y]$
$\boldsymbol{E[X-Y]} = E[X] - E[Y]$
Variance
$\boldsymbol{Var(X)} = E[(X-\mu)^2] = E[X^2] - \mu^2$
$\boldsymbol{Var(nX)} = E[(nX - n\mu)^2] = E[n^2(X-\mu)^2] = n^2 Var(X)$
s.d. $\sigma = \sqrt{Var(X)}$
Covariance and correlation
$\boldsymbol{Cov(X,Y)} = E[XY] - \mu_X \mu_Y$
$\boldsymbol{Cov(aX+b, cY+d)} = ac \cdot Cov(X,Y)$
$\boldsymbol{Cov(X+Y, Z)} = Cov(X,Z) + Cov(Y,Z)$
$\boldsymbol{Cov(X,X)} = Var(X)$
For 2 **dependent** variables X and Y, $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$
For 2 **independent** variables X and Y, $Cov(X,Y) = 0$
$\boldsymbol{Cor(X,Y)} = \rho = Cov(X,Y)/\sigma_X \sigma_Y$
Discrete r.v.
Probability mass function (PMF): $p_X(x) = P(X = x)$ for all values of x
Cumulative distribution function (CDF): $F_X(x) = P(X \leq x)$ for all values of x
Continuous r.v.
Probability density function (PDF): $f(x)$
Cumulative distribution function (CDF): $F(x)$
Distributions
Bernoulli: $X \sim Ber(p), E[X] = p, Var(X) = p(1-p)$
Binomial: $X \sim Bin(n,p), E[X] = np, Var(X) = np(1-p)$
Geometric: $X \sim Geo(p), E[X] = \frac{1}{p}, Var(X) = \frac{1-p}{p^2}$
Uniform: $X \sim U(a,b), E[X] = \frac{a+b}{2}, Var(X) = \frac{1}{12}(b-a)^2$
Exponential: $F(x) = 1 - e^{-\lambda x}, f(x) = \lambda e^{-\lambda x}\ for\ x \geq 0$ $E[X] = 1/\lambda, Var(X) = 1/\lambda^2$
Normal: $X \sim N(\mu, \sigma^2)$
Central Limit Theorem: For any r.v. where n is large, sum (S_n) and average ($\bar{X}_n$) are approximately normal. $S_n \simeq N(n\mu, n\sigma^2)$ and $\bar{X}_n \simeq N(\mu, \frac{\sigma^2}{n})$
Linear Regression
Least squares method: Ensure sum of squared deviations is minimised $y = \beta_1 x + \beta_0$
Approximations: $\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}, \widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\,\bar{x}$
$S_{xx} = \sum(x_i - \bar{x})^2$ $S_{yy} = \sum(y_i - \bar{y})^2$ $S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y})$
Sum of squared residuals (SSR) = $\sum(y_i - \widehat{y_i})^2$, where $\widehat{y_i}$is the approximated value of y_i

Goodness of fit $r^2 = 1 - \frac{SSR}{S_{yy}} = \frac{S_{xy}^2}{S_{xx}S_{yy}}$

Mean Absolute Error (MAE) = $\frac{1}{k}\sum|\widehat{y_i} - y_i|$

Root Mean Square Error (RMSE) = $\sqrt{\frac{1}{k}\sum(\widehat{y_i} - y_i)^2}$

Bayesian Inference

Hypothesis ($\mathbf{H}$): A statement we wish to accept or reject

Prior $\underline{\mathbf{P(H)}}$: What we believe about the hypothesis without any evidence

Likelihood $\mathbf{L(H|E)} = P(E|H)$: Likelihood of the hypothesis

Posterior $\mathbf{P(E|H)}$: Updated belief after seeing evidence

Conditional independence:
$p(x|y,\theta) = p(x|\theta)$
$p(y|x,\theta) = p(y|\theta)$
$p(x,y|\theta) = p(x|\theta)p(y|\theta)$, where $\theta$is the parameter

Updating normal distributions:
If n datapoints $x_1, x_2, \ldots, x_n$ are drawn from N(θ, σ^2), where θ^2 is known,
$\sigma_{post}^2 = 1/(\frac{1}{\sigma_{prior}^2} + \frac{n}{\sigma^2})$
$\mu_{post} = \frac{\sigma_{post}^2}{\sigma_{prior}^2}\mu_{prior} + \frac{n\sigma_{post}^2}{\sigma^2}\bar{x}$
Note also that $\frac{\sigma_{post}^2}{\sigma_{prior}^2} + \frac{n\sigma_{post}^2}{\sigma^2} = 1$, and $\sigma_{post}^2 < \sigma_{prior}^2$

Frequentist Inference

Type I error: Reject H_0 when H_0 is true
Type II error: Do not reject H_0 when H_1 is true

Significance level ($\boldsymbol{\alpha}$) = $P(\text{Type I error})$

Power of a test = $1 - P(\text{Type II error})$ – note that this requires the knowledge of the distribution of our random variable under H_1 as well

Variance approximation: if σ^2 is unknown, we approximate it as $s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$
(i.e. 1/(n-1) * sample variance)

Conversion to T-distribution: Given $X \sim N(\mu_0, \sigma^2)$ with unknown σ^2 ,

$$T = \frac{X - \mu_0}{\frac{\sqrt{s^2}}{\sqrt{n}}} \sim t(n-1)$$

t-test for 2 samples: take pooled sample variance (s_p^2) as

$$s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{(n-1) + (m-1)}(\frac{1}{n} + \frac{1}{m})$$

Maximum Likelihood Estimation

MLE: Consider data x drawn from some distribution with an unknown parameter p. The MLE estimation of p is the value of θ^* that maximises the likelihood $p(x|p)$

$$\theta^* = \arg\max_{\theta} P(x|p = \theta)$$

Geometric and Exponential: $\theta^* = \frac{occurences}{total\ time} = \frac{1}{average\ time\ per\ occurence}$

Bernoulli and Binomial: $\theta^* = \frac{successes}{total}$

Uniform: $\hat{b} = \max(x_1, \ldots, x_n)$, $\hat{a} = \min(x_1, \ldots, x_n)$

Normal: Sample mean and sample variance

Confidence Interval: Consider data drawn from some distribution with an unknown, fixed value θ. The interval estimator $[\widehat{\theta_L}, \widehat{\theta_U}]$ is called a confidence interval if $P(\widehat{\theta_L} \le \theta \le \widehat{\theta_H}) = 1 - \alpha$, where $1 - \alpha$ is the confidence level

Bias of point estimators: $\hat{\theta}$ is an unbiased estimator if $E[\hat{\theta}] = \theta$, and biased otherwise

Classification

Naïve Bayes': Assume that all features are mutually independent.

$$p(x_{i,1}, x_{i,2}, \ldots, x_{i,m}|y_i) = \prod p(x_{i,j}|y_i)$$

Laplacian smoothing: When the likelihood of any feature is 0, add 1 for each class and feature value

	Prediction (1/0)

<table>
<tr><td rowspan="2">Ground truth (1/0)</td><td>True Positive (TP)</td><td>False Negative (FN)</td></tr>
<tr><td>False Positive (FP)</td><td>True Negatve (TN)</td></tr>
<tr><td colspan="3">Accuracy: $\frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$</td></tr>
<tr><td colspan="3">Precision: $\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$</td></tr>
<tr><td colspan="3">Recall: $\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$</td></tr>
<tr><td colspan="3">F-score: Harmonic mean of precision and recall $F_1 = \frac{2}{\mathrm{recall}^{-1}+\mathrm{precision}^{-1}} = 2\frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision}+\mathrm{recall}} = \frac{\mathrm{TP}}{\mathrm{TP}+\frac{1}{2}(\mathrm{FP}+\mathrm{FN})}$</td></tr>
</table>