Results by Module

Last modified: 2026-11-19

Every result in the notes, listed by module and page in the order the notes present them. Each row numbers the result within its page and across the whole list, so the last overall number counts every result. The list is built from the notes themselves, so it cannot fall out of step with them. The definitions are listed the same way.

M0: Prerequisites

Probability

Table 1: The results on Probability, in page order.

On page	Overall	Result
1	1	The chain rule of probability

M2: Linear Regression

Linear Regression

Table 2: The results on Linear Regression, in page order.

On page	Overall	Result
1	2	The padded 1 absorbs the intercept
2	3	The least squares objective is a convex quadratic
3	4	A minimizer is a stationary point
4	5	A stationary point of a convex function is a global minimizer
5	6	Gradient of the squared error at one point
6	7	Gradient of the squared error over all points
7	8	Sums of design vectors as matrix products
8	9	Gradient of the least-squares objective
9	10	The two cross terms of the squared error are equal
10	11	The normal equations
11	12	Shortcut formulas for the centered sums

On page	Overall	Result
12	13	The simple linear regression slope is the sample covariance over the sample variance
13	14	The two simple regression slopes multiply to the squared correlation
14	15	The hat matrix is an orthogonal projection
15	16	Bounds on the leverage
16	17	The mean leverage
17	18	Least squares as maximum likelihood estimation
18	19	The coefficient of determination is negative when the model does worse than the mean
19	20	Reducible error and residual variance decomposition
20	21	Normal equations for expanded features
21	22	The number of polynomial features
22	23	Degree-two features are the entries of a self outer product

Interactions, Inference and Diagnostics

Table 3: The results on Interactions, Inference and Diagnostics, in page order.

On page	Overall	Result
1	24	A model with its main effects is unchanged by shifting a feature
2	25	Shifting a feature can create a dropped main effect
3	26	Mean and covariance of the least squares estimate
4	27	The standardized coefficient has a t distribution
5	28	The coefficient interval has probability 0.95 of containing the coefficient
6	29	The variance of a prediction
7	30	The variance of a prediction error
8	31	The prediction interval is wider than the confidence interval

Multivariate-Outcome Linear Regression

Table 4: The results on Multivariate-Outcome Linear Regression, in page order.

On page	Overall	Result
1	32	The Frobenius objective is a sum over columns
2	33	Multivariate normal equations

M3: ML Fundamentals

A First Model, End to End

Table 5: The results on A First Model, End to End, in page order.

On page	Overall	Result
1	34	Under squared-error loss, the empirical risk is the sample mean squared error (sMSE)

Generalization and Cross-Validation

Table 6: The results on Generalization and Cross-Validation, in page order.

On page	Overall	Result
1	35	Hoeffding's inequality for a held-out set

Bias, Variance and Residual Variance

Table 7: The results on Bias, Variance and Residual Variance, in page order.

On page	Overall	Result
1	36	The noise has population mean zero
2	37	The bias as the average fit minus the true mean function
3	38	The variance equals the variance of the estimation error
4	39	Separating the noise from the expected error
5	40	Separating the bias from the model's own error
6	41	Bias–variance decomposition

M4: Classification and Logistic Regression

Logistic Regression

Table 8: The results on Logistic Regression, in page order.

On page	Overall	Result
1	42	Symmetry of the logistic function
2	43	Derivative of the logistic function
3	44	Relabeling the classes negates the coefficients
4	45	Derivative of the logistic function
5	46	The logistic function is steepest at zero
6	47	The log odds and the logistic function are inverses
7	48	The slope at a fixed probability does not depend on the intercept
8	49	The slope at probability one half is a quarter of the slope coefficient
9	50	The odds multiplier of a coefficient
10	51	The odds ratio of a coefficient
11	52	The logistic regression boundary is a hyperplane

Gradient Descent

Table 9: The results on Gradient Descent, in page order.

On page	Overall	Result
1	53	Scaling the cost rescales the equivalent learning rate
2	54	The cost is a quadratic form in the error plus the minimum cost
3	55	One gradient-descent update multiplies the error by a fixed matrix
4	56	Gradient descent on the residual sum of squares converges exactly when the step is below one over the largest eigenvalue
5	57	The best constant step size for least squares
6	58	The cost along the negative gradient is a parabola
7	59	Consecutive exact-line-search gradients are orthogonal
8	60	The stochastic gradient step is unbiased
9	61	The minibatch step is unbiased
10	62	The minibatch variance is the single-point variance over the batch size

Fitting Logistic Regression

Table 10: The results on Fitting Logistic Regression, in page order.

On page	Overall	Result
1	63	Probability mass of one Bernoulli outcome
2	64	Likelihood of conditionally independent Bernoulli outcomes
3	65	Log-likelihood of conditionally independent Bernoulli outcomes
4	66	Binary form of the log loss
5	67	The cross-entropy loss equals the log loss
6	68	The total log loss is the negative log-likelihood
7	69	Minimizing the log loss is maximum likelihood estimation
8	70	The log loss depends on the margin alone
9	71	Cross-entropy is smallest at the true distribution
10	72	The range of the entropy
11	73	Cross-entropy of a classifier from the probabilities of the observed labels
12	74	Cross-entropy of a classifier is nonnegative
13	75	Changing the base of cross-entropy rescales it
14	76	Derivative of the log loss with respect to the log-odds
15	77	Gradient of the total log loss
16	78	Hessian of the total log loss
17	79	The Hessian of the total log loss has bounded eigenvalues
18	80	A step size that is safe for every coefficient vector
19	81	The residuals are orthogonal to every column of the design matrix at a minimizer
20	82	The intercept of a balanced two-group fit
21	83	The logistic log-likelihood is strictly concave
22	84	The Newton update for logistic regression
23	85	A Newton step is a weighted least squares fit
24	86	The Newton error constant of the worked fit

Predictions and Inference for Logistic Regression

Table 11: The results on Predictions and Inference for Logistic Regression, in page order.

On page	Overall	Result
1	87	Separated data have no maximum likelihood estimate

Multiclass Logistic Regression

Table 12: The results on Multiclass Logistic Regression, in page order.

On page	Overall	Result
1	88	The two-class multinomial model is logistic regression
2	89	Gradient of the multinomial log-likelihood

Discriminant Analysis

Table 13: The results on Discriminant Analysis, in page order.

On page	Overall	Result
1	90	The posterior from a prior and class densities
2	91	The linear discriminant function
3	92	LDA implies a logistic model
4	93	The quadratic discriminant function
5	94	Quadratic discriminant boundaries are quadric surfaces
6	95	Parameter counts of the classifiers

M5: Performance Metrics

Metrics

Table 14: The results on Metrics, in page order.

On page	Overall	Result
1	96	Micro-averaged precision and recall equal accuracy
2	97	A score independent of the label has ROC AUC one half

M6: Regularization and Model Selection

Subset Selection

Table 15: The results on Subset Selection, in page order.

On page	Overall	Result
1	98	Number of models that best subset selection fits

Regularization

Table 16: The results on Regularization, in page order.

On page	Overall	Result
1	99	A gradient step on the ridge objective is weight decay
2	100	The coordinate update for the lasso
3	101	Ridge and the lasso with orthonormal features
4	102	The coefficients of a dimension reduction regression

Model Selection

Table 17: The results on Model Selection, in page order.

On page	Overall	Result
1	103	A larger model family never has a larger minimum empirical risk
2	104	The variance of the cross-validation estimate in terms of fold correlation

M7: Support Vector Machines and Kernel Methods

Support Vector Machines

Table 18: The results on Support Vector Machines, in page order.

On page	Overall	Result
1	105	The margin of a point, in one dimension
2	106	The max-margin program has a ray of optima
3	107	Fixing the scale of the max-margin program
4	108	The margin of a point, in p dimensions
5	109	The optimal weight vector is a weighted sum of the training inputs
6	110	Gradient descent from zero stays in the span of the training inputs

On page	Overall	Result
7	111	Only points on the margin carry weight
8	112	Recovering $\vec{\beta}$ and β_0 from the dual
9	113	The total violation never increases with C
10	114	The margin never increases with C
11	115	The optimal slack is the hinge loss
12	116	The gradient of the C-SVM objective away from its corners
13	117	The decision rule of a kernel SVM

M8: Neural Networks and Backpropagation

Neural Networks

Table 19: The results on Neural Networks, in page order.

On page	Overall	Result
1	118	XOR is not linearly separable
2	119	The product feature separates XOR
3	120	Number of polynomial features
4	121	Parameters of a one-hidden-layer network

Backpropagation

Table 20: The results on Backpropagation, in page order.

On page	Overall	Result
1	122	The gradient of the last layer's weights is an outer product
2	123	The gradient of the first layer's weights is an outer product
3	124	Derivative of the hyperbolic tangent

M10: Clustering

Unsupervised Learning

Table 21: The results on Unsupervised Learning, in page order.

On page	Overall	Result
1	125	K-means never raises the distortion

Optional Reading: Theory

Machine Learning Theory

Table 22: The results on Machine Learning Theory, in page order.

On page	Overall	Result
1	126	The VC generalization bound

Optional Reading: Non-Parametric Models

Splines

Table 23: The results on Splines, in page order.

On page	Overall	Result
1	127	Parameters of a cubic spline

Generalized Additive Models

Table 24: The results on Generalized Additive Models, in page order.

On page	Overall	Result
1	128	The backfitting update for one component

Optional Reading: Further Methods

Gaussian Processes

Table 25: The results on Gaussian Processes, in page order.

On page	Overall	Result
1	129	Conditioning a scalar Gaussian
2	130	Gaussian process regression

Convolutional and Recurrent Networks

Table 26: The results on Convolutional and Recurrent Networks, in page order.

On page	Overall	Result
1	131	Output size of a convolution
2	132	Jacobian of the state of a tanh recurrent layer
3	133	Jacobian of a tanh recurrent layer
4	134	The backward recursion and the shared-weight gradient

Attention and Transformers

Table 27: The results on Attention and Transformers, in page order.

On page	Overall	Result
1	135	Perplexity does not depend on the base