ML Tutorial I

Preamble

$$\int a \cdot f(x) + b \cdot g(x)\, dx = a\int f(x)\, dx + b\int g(x)\, dx$$

Probability

Property of probability distribution "by definition": $\int p(X)\, dX = 1$

Definition of conditional probability: $P(X \mid Y) = \frac{P(X,Y)}{P(Y)}$. Think of this as "zooming in" onto the subspace represented by Y.

Conditional independence: if two variables are conditionally independent, then $P(X,Y) = P(X)P(Y)$

A few tricks:

$$P(X,Y) = P(X \mid Y)P(Y) = P(Y \mid X)P(X)$$

$$P(X \mid Y) = \frac{P(X,Y)}{P(Y)} = \frac{P(Y \mid X)P(X)}{P(Y)} = \frac{P(Y \mid X)P(X)}{\int P(Y \mid X)P(X)\, dX}$$

for a given Y, $P(X \mid Y) = \frac{P(X,Y)}{P(Y)} \propto P(X,Y)$

$$\int p(X,Y)\, dX = \int p(X \mid Y)P(Y)\, dX = P(Y)\int p(X \mid Y)\, dX = P(Y)\cdot 1$$

Expectated value (definition): $E[X] = \int Xp(X)\, dX$

Fact: $E[aX + b] = aE[X] + b$

Proof: $E[aX + b] = \int (aX + b)\, p(X)\, dX = a\int Xp(X)\, dX + b\int p(X)\, dX = aE[X] + b$

Fact: Expectation of a sum of RV (random variables) is the sum of their expectations EVEN WHEN THEY ARE NOT INDEPENDENT!

Proof: $E[X+Y] = \int\int (X+Y)\, p(X,Y)\, dYdX = \int\int Xp(X,Y)\, dYdX + \int\int Yp(X,Y)\, dYdX =$

$$= \int\int Xp(X,Y)\, dYdX + \int\int Yp(X,Y)\, dXdY = \int\int Xp(Y \mid X)\, p(X)\, dYdX + \int\int Yp(X \mid Y)\, p(Y)\, dXdY =$$

$$= \int Xp(X)\int p(Y \mid X)\, dYdX + \int Yp(Y)\int p(X \mid Y)\, dXdY =$$

$$= \int Xp(X)\, \cancel{\int p(Y \mid X)\, dY}\, dX + \int Yp(Y)\, \cancel{\int p(X \mid Y)\, dX}\, dY = E[X] + E[Y]$$

Fact: you generally can't do the same with the product or 2 RVs: $E[X \cdot Y] \neq E[X] \cdot E[Y]$

Fact: for INDEPENDENT RVs, $E[X \cdot Y] = E[X] \cdot E[Y]$

Proof: $E[X \cdot Y] = \int\int XYp(X,Y)\,dYdX = \int\int XYp(X)\,p(Y)\,dYdX =$

$= \int Xp(X)\left[\int Yp(Y)\,dY\right]dX = \int Xp(X)\,E[Y]\,dX = E[X] \cdot E[Y]$

Variance (definition): $Var[X] = E\left[\left(X - E[X]\right)^2\right]$. In plain language, it can be expressed as "The average of the square of the distance of each data point from the mean".

Property: $Var[aX + b] = a^2 Var[X]$

Proof: $Var[aX + b] = E\left[\left(aX + b - E[aX + b]\right)^2\right] = E\left[\left(aX + b - aE[X] - b\right)^2\right] =$

$= E\left[\left[a\left(X - E[X]\right)\right]^2\right] = a^2 E\left[\left(X - E[X]\right)^2\right] = a^2 Var[X]$

Fact: $Var[X] = E\left[X^2\right] - \left[E(X)\right]^2$ -- this fact is very useful when you are given samples from X one batch at a time, as $E[X_1 + X_2] = E[X_1] + E[X_2]$, and $E\left[X_1^2 + X_2^2\right] = E\left[X_1^2\right] + E\left[X_2^2\right]$, but $Var[X_1 + X_2] \neq Var[X_1] + Var[X_2]$ (OK, not really THAT useful)

Proof: $Var[X] = E\left[\left(X - E[X]\right)^2\right] = E\left[X^2 - X \cdot E[X] - E[X] \cdot X + E[X]^2\right] =$

$= E\left[X^2\right] - 2E\left[X \cdot E[X]\right] + \left[E(X)\right]^2 = E\left[X^2\right] - 2\left[E(X)\right]^2 + \left[E(X)\right]^2 =$

$= E\left[X^2\right] - \left[E(X)\right]^2$

Covariance (definition): $Cov(X,Y) = E\left[\left(X - E[X]\right)\left(Y - E[Y]\right)\right]$

Obvious fact: $Cov(X,X) = Var(X)$

Obvious fact: $Cov(X,Y) = Cov(Y,X)$

Fact: for independent RV (and also for uncorrelated ones by definition), $Cov(X,Y) = 0$

Proof for independent RVs: $Cov(X,Y) = E\left[\left(X - E[X]\right)\left(Y - E[Y]\right)\right] =$

$= E\left[\left(X - E[X]\right)\right]E\left[\left(Y - E[Y]\right)\right] = \left(E[X] - E[X]\right)\left(E[Y] - E[Y]\right) = 0$

Fact: $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X,Y)$

Proof : $Var(X + Y) = E\left[\left(X - E[X] + Y - E[Y]\right)^2\right] = E\left[\left[\left(X - E[X]\right) + \left(Y - E[Y]\right)\right]^2\right] =$

$= E\left[\left(X - E[X]\right)^2\right] + E\left[\left(Y - E[Y]\right)^2\right] + 2E\left[\left(X - E[X]\right)\left(Y - E[Y]\right)\right] =$

$= Var(X) + Var(Y) + 2Cov(X,Y)$

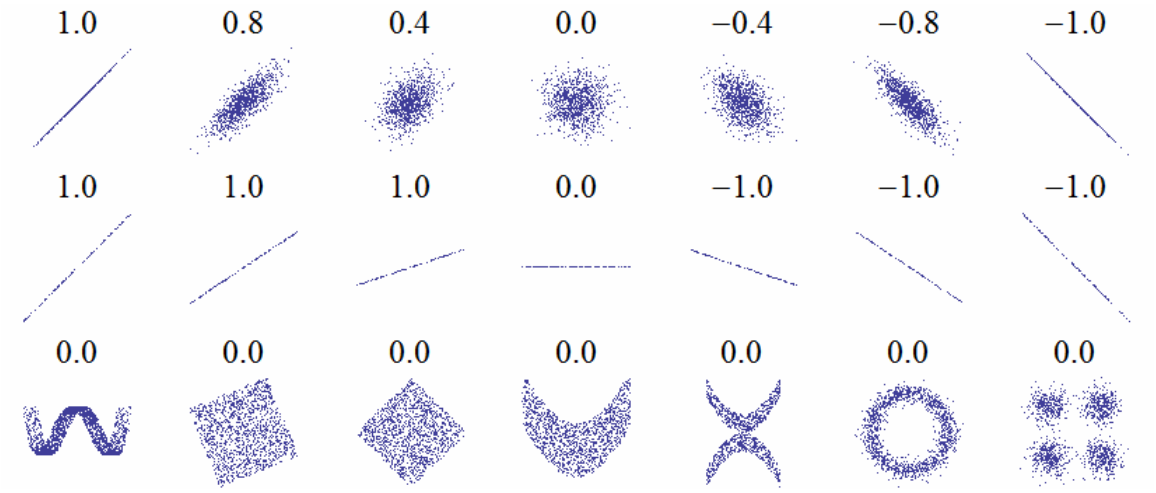
Fact: If X and Y are UNCORRELATED (not necessarily INDEPENDENT), then

$Var(X + Y) = Var(X) + Var(Y)$

Proof: If the covariance of the RVs is 0, then

$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X,Y) = Var(X) + Var(Y)$

Covariance only refers to a LINEAR relationship between the two RVs. Here's an informative picture from Wikipedia (the number shown is correlation, which is proportional to covariance, but we don't need it for machine learning):



A covariance matrix is simply the matrix composed of all of the covariances between each pair of RVs in a vector RV. Which is the perfect segue into…

Linear Algebra

A matrix is a box of numbers. Example: $M = \begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix}$. This matrix has 2 rows and 3 columns. The entry on the 2nd row, 3rd column is $M_{2,3} = f$

Vector (definition): matrix with only one row (row vector) or only one column (column vector). By tradition, vectors are considered to be COLUMN VECTORS, unless otherwise specified.

Transpose of a matrix (definition): Given a matrix $[A]_{ij}$, the matrix $\left[A^T\right]_{ji}$ is its transpose. Basically it flips rows with columns. Transposing a scalar (or a 1x1 matrix) does not change it.

Square matrix (definition): A matrix with an equal number of rows and columns

Main diagonal (definition): the set of elements: $a_{11}, a_{22}, a_{33}, ..., a_{ii}$

Diagonal matrix (definition): a matrix with non-zero entries only on its main diagonal

Identity matrix (definition): A diagonal matrix with all ones on its main diagonal. Notation: **I**
Fact: The identity matrix is the NEUTRAL ELEMENT of matrix multiplication.

Matrix addition (definition): $[A+B]_{ij} = [A]_{ij} + [B]_{ij}$

Matrix subtraction (definition): $[A-B]_{ij} = [A]_{ij} - [B]_{ij}$

Matrix addition and subtraction only makes sense when the two matrices have the same number of rows and columns. Addition is commutative. The neutral element is the matrix of all zeros, also known as the zero matrix, notated with **0**.

Matrix multiplication (definition): $[AB]_{ij} = \sum_k [A]_{ik} [B]_{kj}$.

Matrix multiplication only makes sense when the number of columns of A is the same as the number of rows of B. Matrix multiplication is not commutative. The neutral element is the identity matrix.

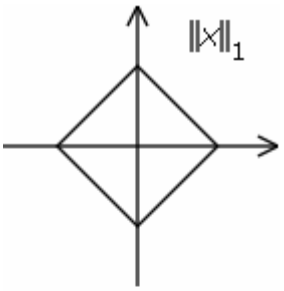
Property: $(AB)^T = B^T A^T$. This comes from putting together the definition of the transpose and that of the matrix multiplication (do it yourself!).

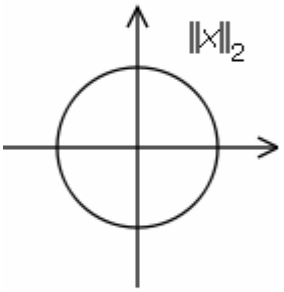
Dot product (definition): The matrix multiplication of a horizontal vector and a vertical vector $x \cdot y = x^T y$. Its result is a scalar.

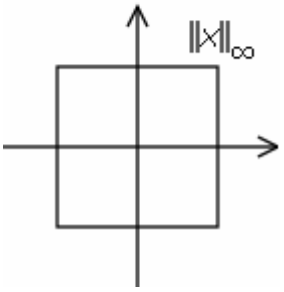
Euclidian norm (or 2-norm): $\|x\| = \sqrt{x^T x} = \left(\sum_i x_i^2 \right)^{\frac{1}{2}}$. This is the Euclidian distance

Normalization (definition): Dividing a vector by its norm – the result is a vector of the same direction and norm 1.

p-norm (definition): $\|x\|_p = \left(\sum_i |x_i|^p \right)^{\frac{1}{p}}$







Property of dot product: $x^T y = \|x\| \|y\| \cos\alpha$, α is the angle between vectors x and y . This is proven via the law of cosines.

Orthogonal (or perpendicular) vectors: vectors whose dot product is 0. This happens when the angle between them is 90°, as that makes its cosine 0 (see formula above).

Outer product (definition): The matrix multiplication of a vertical vector and a horizontal vector $x \otimes y = xy^T$. Its result is a matrix.

Eigenvectors and eigenvalues of a SQUARE matrix (definition): An eigenvector-eigenvalue pair of a square matrix M is formed of the vector **v** and value λ such that $\mathbf{Mv} = \lambda \mathbf{v}$

Symmetric matrix (definition): Matrix that is equal to its own transpose: $M = M^T$
Property: All eigenvectors of a symmetric matrix are orthogonal (except for cases with repeated eigenvalues).
Proof: Let M be a symmetric matrix. Let $\lambda_1 \neq \lambda_2$ be eigenvalues, and v_1, v_2 be the corresponding eigenvectors, so $M\lambda_1 = v_1\lambda_1$ and $M\lambda_2 = v_2\lambda_2$.

$\lambda_1 \left(v_1^T v_2 \right) = \left(\lambda_1 v_1 \right)^T v_2 = \left(M v_1 \right)^T v_2 = v_1^T M^T v_2 \left(\text{key point!} \right) = v_1^T M v_2 = v_1^T \lambda_2 v_2 = \lambda_2 \left(v_1^T v_2 \right).$

Since $\lambda_1 \left(v_1^T v_2 \right) = \lambda_2 \left(v_1^T v_2 \right)$, $\left(\lambda_2 - \lambda_1 \right) \left(v_1^T v_2 \right) = 0$, but $\lambda_2 - \lambda_1 \neq 0$, therefore $v_1^T v_2 = 0$, which means that the vectors are orthogonal

Spectral decomposition (or eigendecomposition): This works for both symmetric and non-symmetric matrices. However in this course you'll only need to apply it to symmetric matrices.

If we construct a matrix **L** out of column eigenvectors, and a diagonal matrix **D**, with the eigenvalues on its diagonal, then we can express the symmetric matrix **M**:

$\mathbf{M} = \mathbf{L}\mathbf{D}\mathbf{L}^{\mathbf{T}}$

(I've just revised this one – I had gotten it wrong – thanks to Francisco for pointing it out)

Quadratic form (definition – term from statistics): the scalar quantity $x^T M x$

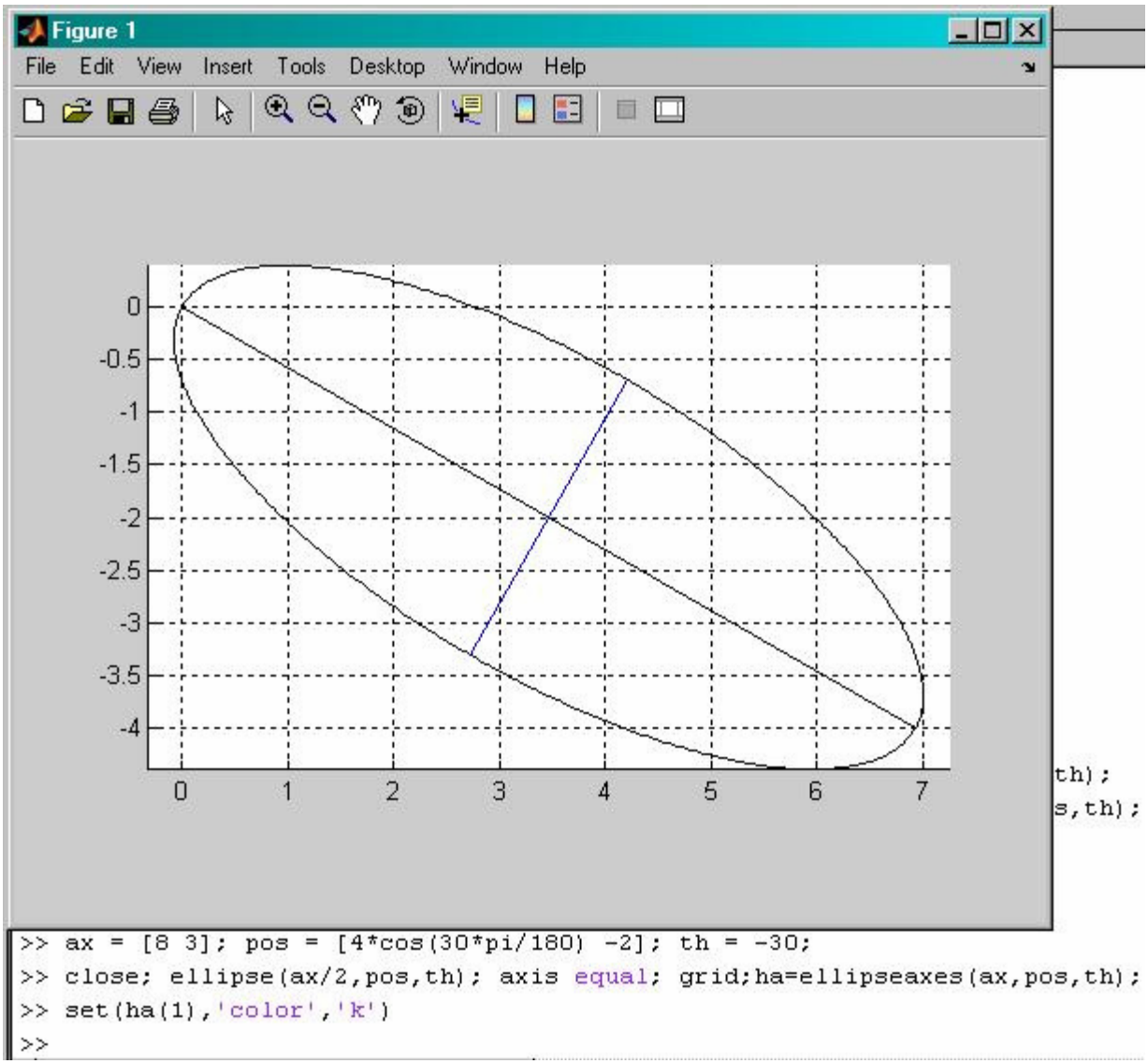
Positive semidefinite matrix (definition): square and SYMMETRIC matrix with the property that for any x, the quadratic form $x^T M x \geq 0$

Property: All eigenvalues of a positive semidefinite matrix are non-negative.

Property: Because it's symmetric, all eigenvectors of a positive semidefinite matrix are orthogonal.

! Time to digest all of the above !

One may visualize a positive-semidefinite matrix or the corresponding quadratic form through an n-dimensional ellipsoid centered at the origin. This ellipsoid is the set of all points of the quadratic form $x^T M x$ with the property $\|x\| = 1$ The direction of the axes of the ellipse is given by the eigenvectors of the matrix, while the relative length of each axis is given by the corresponding eigenvalue. The identity matrix produces an n-sphere. A diagonal matrix produces an ellipsoid that has the cartesian axes for its axes. [More to come].



The (multivariate) normal probability distribution

If we take a quadratic form (multiplied by -1/2) and put it inside the exponential function, we get the bell-shaped normal (also known as Gaussian) probability distribution:

$$N(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) \propto \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right].$$

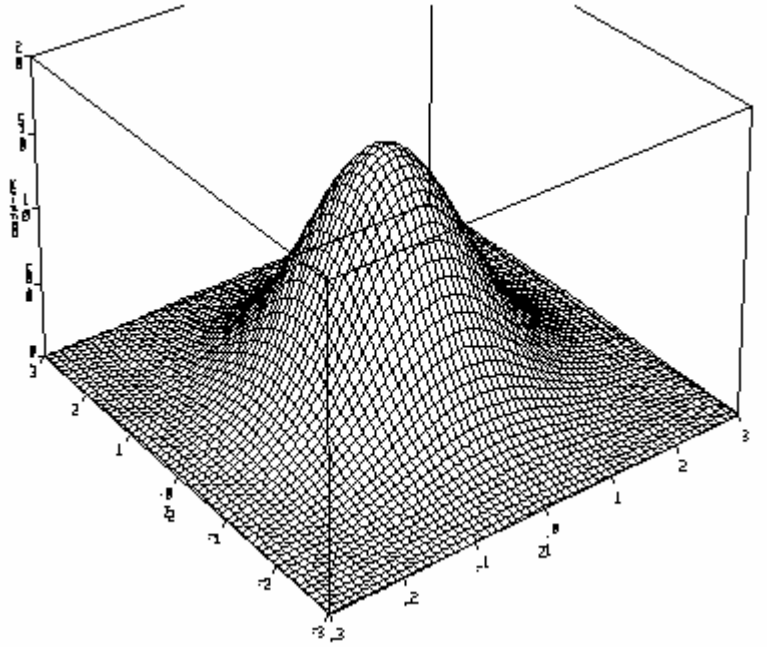
This distribution is parametrized by $\boldsymbol{\mu}$ (its mean), and $\boldsymbol{\Sigma}$ (its covariance).

This distribution has some nice properties. Let X ~ Normal$(\boldsymbol{\mu},\boldsymbol{\Sigma})$, and Y ~ Normal . Then:

- $E[X]=\boldsymbol{\mu}$, $Cov[X]=\Sigma$
- $X+Y$ ~ Normal, $X-Y$ ~ Normal
- $p(X)\cdot p(Y)$ ~ Multivariate normal
- $X \mid Y$ ~ Normal

- FourierTransform$[X]$ has the same Gaussian form (exponential of -1/2*quadratic form)
- The Central Limit Theorem states: Let $X_1, X_2, X_3, ..., X_n$ be a sequence of n independent and identically distributed (i.i.d.) random variables each having finite values of expectation μ and variance $\sigma^2 > 0$. The central limit theorem states that as the sample size n increases, the distribution of the sample *average* of these random variables approaches the normal distribution with a mean μ and variance σ^2/n irrespective of the shape of the original distribution.

To compute the normalizer of this distribution we're going to need some more linear algebra…



Linear algebra continued

Check out: http://www.cs.toronto.edu/~roweis/notes/matrixid.pdf

Trace of a square matrix (definition): the sum of the diagonal entries: $Tr[A] = \sum_k [A]_{kk}$

Property: The trace of a matrix is equal to the sum of its eigenvalues.
Property: $Tr[ABC] = Tr[BCA] = Tr[CAB]$, but $Tr[ABC] \neq Tr[BAC]$. This comes from putting together the definitions for trace and for matrix multiplication.

Inverse of a SQUARE matrix (definition): Given a square matrix M, its inverse is the matrix M^{-1}, such that $MM^{-1} = M^{-1}M = I$.
Fact: The inverse does not exist if ANY of the eigenvalues is zero.

Determinant of a square matrix (definition):
The determinant of a square matrix $A = [\mathbf{a}_1, ..., \mathbf{a}_n]$ is the unique real-valued function of n vectors in $\mathbb{R}^n$ with the following properties:

1. *Multilinearity:* det A is linear with respect to each of its arguments
2. *Antisymmetry:* Exchanging any two arguments changes its sign
3. *Normalization:* The determinant of the identity matrix is 1

One can prove both the existence and uniqueness of the determinant, but my textbook says that the proof is best left for a rainy day (it must not have been written in London – haha).

Property: The determinant of a matrix is equal to the product of its eigenvalues.
Property: If we construct an *n*-parallelotope from the row vectors of the matrix, the determinant is the *n*-parallelotope's volume.
Relationship between inverse and determinant: $\mathbf{X}^{-1} = |\mathbf{X}|^{-1} \operatorname{adj}(\mathbf{X})$

Precision matrix (definition): the inverse of the covariance matrix

Multivariate Calculus

A few formulas from univariate calculus first:

$$\frac{d}{dx}\log x = \frac{1}{x}$$
$$\frac{d}{dx}(u \cdot v) = u\frac{dv}{dx} + v\frac{du}{dx}$$
$$\frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx}$$

We are going to take derivatives of scalars with respect to vectors, derivatives of vectors with respect to scalars, derivatives of vectors with respect to vectors, derivatives of scalars with respect to matrices etc. To make sense out of these, we are going to compute them at the univariate calculus level, and then put them together into a vector or matrix.

$$\frac{d(\mathbf{Ax})}{d\mathbf{x}} \to \frac{d[\mathbf{Ax}]_i}{dx_j} = \frac{d}{dx_j}\sum_k A_{ik}x_k = \frac{d}{dx_j}A_{ij}x_j + \cancel{\sum_{k\neq j} A_{ik}x_k} = A_{ij} \to \frac{d(\mathbf{Ax})}{d\mathbf{x}} = \mathbf{A}^T \text{ (or } \mathbf{A}\text{, depending on convention)}$$

$$\frac{d(\mathbf{x^T Ax})}{d\mathbf{x}} \to \frac{d(\mathbf{x^T Ax})}{dx_j} = \frac{d}{dx_j}\sum_i\sum_k x_i A_{ik} x_k = \frac{d}{dx_j}\left(x_j A_{jj} x_j + x_j \sum_{k\neq j} A_{jk}x_k + \left(\sum_{i\neq j} x_i A_{ij}\right) x_j + \cancel{\sum_{i\neq j}\sum_{k\neq j} x_i A_{ik} x_k} \right) =$$
$$= \frac{d}{dx_j}\left(x_j A_{jj} x_j + x_j \sum_{k\neq j} A_{jk}x_k + x_j \sum_{k\neq j} A_{kj}x_k \right) = 2x_j A_{jj} + \sum_{k\neq j} A_{jk}x_k + \sum_{k\neq j} A_{kj}x_k =$$
$$= \sum_k A_{jk}x_k + \sum_k A_{kj}x_k = [\mathbf{Ax}]_j + [\mathbf{A^T x}]_j = \left[(\mathbf{A} + \mathbf{A^T})\mathbf{x}\right]_j \to \frac{d(\mathbf{x^T Ax})}{d\mathbf{x}} = (\mathbf{A} + \mathbf{A^T})\mathbf{x}$$

$$\frac{d|\mathbf{X}|}{d\mathbf{X}} \rightarrow \frac{d|\mathbf{X}|}{d\mathbf{X}_{ij}} = \frac{d}{d\mathbf{X}_{ij}} \sum_k X_{qk} \left[\text{adj}^T (X) \right]_{qk} = \sum_k \frac{d}{d\mathbf{X}_{ij}} \left(X_{ik} \left[\text{adj}^T (X) \right]_{ik} \right) =$$

$$\sum_k \left(\frac{d}{d\mathbf{X}_{ij}} X_{ik} \right) \left[\text{adj}^T (X) \right]_{ik} + \left(\frac{d}{d\mathbf{X}_{ij}} \left[\text{adj}^T (X) \right]_{ik} \right) X_{ik}, \text{but } \frac{d}{d\mathbf{X}_{ij}} \left[\text{adj}^T (X) \right]_{ik} = 0, \text{ so}$$

$$\frac{d|\mathbf{X}|}{d\mathbf{X}_{ij}} = \sum_k \left(\frac{d\mathbf{X}_{ik}}{d\mathbf{X}_{ij}} \right) \left[\text{adj}^T (X) \right]_{ik}, \text{ but } \frac{d\mathbf{X}_{ik}}{d\mathbf{X}_{ij}} = \delta_{jk}, \text{ so}$$

$$\frac{d|\mathbf{X}|}{d\mathbf{X}_{ij}} = \left[\text{adj}^T (X) \right]_{ij}. \text{ Since } \mathbf{X}^{-1} = |\mathbf{X}|^{-1} \text{adj}(X) \rightarrow \text{adj}(X) = \mathbf{X}^{-1} |\mathbf{X}| \rightarrow$$

$$\frac{d|\mathbf{X}|}{d\mathbf{X}} = |\mathbf{X}| \left(\mathbf{X}^{-1} \right)^{\mathbf{T}}$$

$$\frac{d \log|\mathbf{X}|}{d\mathbf{X}} = \frac{d \log|\mathbf{X}|}{d|\mathbf{X}|} \frac{d|\mathbf{X}|}{d\mathbf{X}} = |\mathbf{X}|^{-1} |\mathbf{X}| \left(\mathbf{X}^{-1} \right)^{\mathbf{T}} = \left(\mathbf{X}^{-1} \right)^{\mathbf{T}}$$

Reference: see Jacobi's formula and Adjugate matrix on Wikipedia

Other

$$p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\exp\left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^{\mathbf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right]}{\int \exp\left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^{\mathbf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right] d\mathbf{x}} = (2\pi)^{-N/2} |\Sigma|^{-1/2} \exp\left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^{\mathbf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right]$$

How do we compute this? Start with

$$\left(\int \exp\left[-\frac{1}{2}x^2/\sigma^2\right]dx\right)^2 = \left(\int \exp\left[-\frac{1}{2}x^2/\sigma^2\right]dx\right)\left(\int \exp\left[-\frac{1}{2}y^2/\sigma^2\right]dy\right) =$$

$$= \int\int \exp\left[-\frac{1}{2}x^2/\sigma^2\right]\exp\left[-\frac{1}{2}y^2/\sigma^2\right]dydx = \int\int \exp\left[-\frac{1}{2}x^2/\sigma^2 - \frac{1}{2}y^2/\sigma^2\right]dydx =$$

$$= \int\int \exp\left[-\frac{1}{2}\left(x^2+y^2\right)/\sigma^2\right]dydx \rightarrow \begin{matrix} x = r\cos\theta \\ y = r\sin\theta \end{matrix} \rightarrow$$

$$\rightarrow |J| = \left|\frac{\partial(x,y)}{\partial(r,\theta)}\right| = \begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{vmatrix} = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = r\cos^2\theta + r\sin^2\theta = r \rightarrow$$

$$\ldots\int_0^{2\pi}\int_0^{\infty} \exp\left[-\frac{1}{2}\left(r^2\cos^2\theta + r^2\sin^2\theta\right)/\sigma^2\right]|J|drd\theta = \int_0^{2\pi}\int_0^{\infty} \exp\left[-\frac{1}{2}r^2/\sigma^2\right]rdrd\theta =$$

$$= 2\pi\int_0^{\infty} r\exp\left[-\frac{1}{2}r^2/\sigma^2\right]dr = 2\pi\left(-\sigma^2\exp\left[-\frac{1}{2}r^2/\sigma^2\right]\right)\Big|_0^{\infty} = 2\pi\sigma^2$$

$$\left(\int \exp\left[-\frac{1}{2}x^2/\sigma^2\right]dx\right)^2 = 2\pi\sigma^2 \rightarrow \int \exp\left[-\frac{1}{2}x^2/\sigma^2\right]dx = \sqrt{2\pi\sigma^2}$$

$$N\left(x;\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2}\left(x-\mu\right)^2/\sigma^2\right]$$

What about the multivariate case? We can use the spectral theorem to construct RVs that would express the multivariate gaussian through independent RVs. Each of these new independent RVs would have a variance corresponding to an eigenvalue of the variance matrix. The product of all of the variances is equal to the product of all of the eigenvalues which is equal to the determinant of the covariance matrix.