Outline

Multivariate Parameters - I

- Mean vector:
$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^T$$
- Covariance of x_i and x_j:
$$\sigma_{ij} = \mathrm{Cov}(x_i, x_j) = \mathbb{E}[(x_i - \mu_i)(x_j - \mu_j)] = \mathbb{E}[x_i x_j] - \mu_i \mu_j$$
Typically the features are correlated, or else there will not be a need for multivariate analysis.
- The x_i and x_j are called uncorrelated if $\sigma_{ij} = \mathbb{E}[x_i x_j] - \mu_i \mu_j = 0$.
- The covariance between two random variables measures the degree to which they are (linearly) related.
- Variance of x_i:
$$\sigma_i^2 = \mathbb{E}[(x_i - \mu_i)^2]$$
- Note that:
$$\sigma_{ij} = \sigma_{ji} \qquad \sigma_{ii} = \sigma_i^2$$

Multivariate Parameters - II

- Covariance matrix:

$$\boldsymbol{\Sigma} \equiv \mathrm{Cov}(\mathbf{x}) = \mathbb{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T] = \mathbb{E}[\mathbf{x}\mathbf{x}^T] - \boldsymbol{\mu}\boldsymbol{\mu}^T = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & & & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{bmatrix}$$

- Correlation between x_i and x_j:

$$\rho_{ij} \equiv \mathrm{Corr}(x_i, x_j) = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

  The correlation (a.k.a. Pearson correlation coefficient) between x_i and x_j is in $[-1, +1]$, making it easier to interpret than the covariance.
  - $\rho_{ij} \neq 0$: two variables x_i and x_j are related in a linear way
- Dependence vs. correlation:

$$x_i \text{ and } x_j \text{ are independent } \underset{\not\Leftarrow}{\overset{\Rightarrow}{}} \sigma_{ij} = \rho_{ij} = 0$$

Parameter Estimation

- Sample mean:

$$\mathbf{m} = \frac{1}{N}\sum_{t=1}^{N}\mathbf{x}^t$$

- Sample covariance matrix:

$$\mathbf{S} = [s_{ij}]_{i,j=1}^{d} = \frac{1}{N}\sum_{t=1}^{N}(\mathbf{x}^t - \mathbf{m})(\mathbf{x}^t - \mathbf{m})^T$$

  where $s_{ii} = s_i^2$

- Sample correlation matrix:

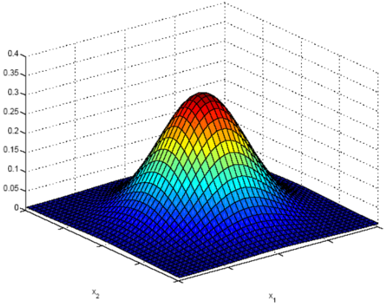
$$\mathbf{R} = [r_{ij}]_{i,j=1}^{d} \quad \text{where} \quad r_{ij} = \frac{s_{ij}}{s_i s_j}$$

Estimation of Missing Values

- What to do if the values of certain variables in some instances are missing?
- Discarding the instances: not a good idea if the sample is small and since the non-missing entries do contain information.
- Imputation: filling in the missing entries
  - Mean imputation: using the most likely value (e.g., mean or mode)
  - Imputation by regression: predicting the missing values based on the regression approach
  - Matrix factorization: using low-rank matrices as factors for matrix completion.

Outline

Multivariate Normal Distribution - I



$$\mathbf{x} \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

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

Multivariate Normal Distribution - II

- Multivariate generalization of univariate normal distribution.
- Multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $d \times 1$ mean vector $\boldsymbol{\mu}$ and $d \times d$ covariance matrix $\boldsymbol{\Sigma}$.
- Probability density function:

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

- Log likelihood:

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \mathcal{X}) = -\frac{Nd}{2}\log(2\pi) - \frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{t=1}^{N}(\mathbf{x}^t - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}^t - \boldsymbol{\mu})$$

- Given sample $\mathcal{X} = \{x^t\}_{t=1}^{N}$, ML estimates:

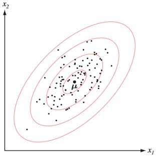
$$\mathbf{m} = \frac{1}{N}\sum_{t=1}^{N}\boldsymbol{x}^t \qquad \mathbf{S} = \frac{1}{N}\sum_{t=1}^{N}(\mathbf{x}^t - \mathbf{m})(\mathbf{x}^t - \mathbf{m})^T$$

Multivariate Normal Distribution - III

- Mahalanobis distance measures the distance from $\mathbf{x}$ to $\boldsymbol{\mu}$ in terms of $\boldsymbol{\Sigma}$ (normalized for differences in variance and covariance):

$$(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

- $(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2$ is the d-dimensional hyperellipsoid centered at $\boldsymbol{\mu}$. Its shape and orientation are defined by $\boldsymbol{\Sigma}$.



- Euclidean distance is a special case of Mahalanobis distance when $\boldsymbol{\Sigma} = s^2\mathbf{I}$; the hyperellipsoid degenerates into a hypersphere.

Bivariate Normal Distribution - I

- Multivariate normal distribution with $d = 2$.
- Covariance matrix:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_2\sigma_1 & \sigma_2^2 \end{bmatrix}$$

- Joint density:

$$p(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}(z_1^2 - 2\rho z_1 z_2 + z_2^2)\right]$$

where

$$z_i = \frac{x_i - \mu_i}{\sigma_i} \text{ (z-normalization)}$$

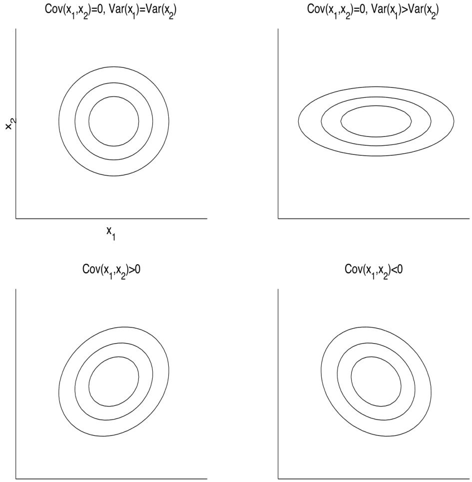
Bivariate Normal Distribution - II

- for $|\rho| < 1$, the equation of an ellipse

$$z_1^2 - 2\rho z_1 z_2 + z_2^2 = c^2$$

  - if $\rho > 0$, the major axis of the ellipse has a positive slope
  - if $\rho < 0$, the major axis of the ellipse has a negative slope
  - If $\rho = 0$, the two variables are independent, the cross-term disappears, and we get a product of two univariate densities.
- If $\rho = \pm 1$, the two variables are linearly related, the observations are effectively one-dimensional, and one of the two variables can be disposed of.

Isoprobability Contour Plot of Bivariate Normal



Independent Inputs

- If x_i are independent, the off-diagonal entries σ_{ij}, $i \neq j$ of $\mathbf{\Sigma}$ are 0. The joint density becomes:

$$p(\mathbf{x}) = \prod_{i=1}^{d} p_i(x_i) = \frac{1}{(2\pi)^{d/2} \prod_{i=1}^{d} \sigma_i} \exp\left[-\frac{1}{2} \sum_{i=1}^{d} \left(\frac{x_i - \mu_i}{\sigma_i}\right)^2\right]$$

  Mahalanobis distance reduces to weighted Euclidean distance (with weightings $1/\sigma_i$).
- It further reduces to Euclidean distance if all variances σ_i^2 are equal.

Outline

Parametric Classification

- In Bayes' decision rule for classification, the discriminant function for of class C_i is

$$p(\mathbf{x} \mid C_i)P(C_i) \quad \text{or} \quad \log\left[p(\mathbf{x} \mid C_i)P(C_i)\right]$$

- Class-conditional densities $p(\mathbf{x} \mid C_i) \sim \mathcal{N}_d(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$:

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

- Discriminant functions:

$$\begin{aligned} g_i(\mathbf{x}) &= \log p(\mathbf{x} \mid C_i) + \log P(C_i) \\ &= -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) + \log P(C_i) \end{aligned}$$

Estimation of Parameters

- Given a training sample for $K \geq 2$ classes, $\mathcal{X} = \{(\mathbf{x}^t, \mathbf{r}^t)\}_{t=1}^N$, where $r_i^t = 1$ if $\mathbf{x}^t \in C_i$ and 0 otherwise, parameters can be estimated separately for each class.
- Parameter estimates:

$$\hat{P}(C_i) = \frac{1}{N} \sum_t r_i^t$$

$$\mathbf{m}_i = \frac{\sum_t r_i^t \mathbf{x}^t}{\sum_t r_i^t}$$

$$\mathbf{S}_i = \frac{\sum_t r_i^t (\mathbf{x}^t - \mathbf{m}_i)(\mathbf{x}^t - \mathbf{m}_i)^T}{\sum_t r_i^t}$$

Quadratic Discriminant Functions - I

- The parameter estimates are then plugged into the discriminant functions:

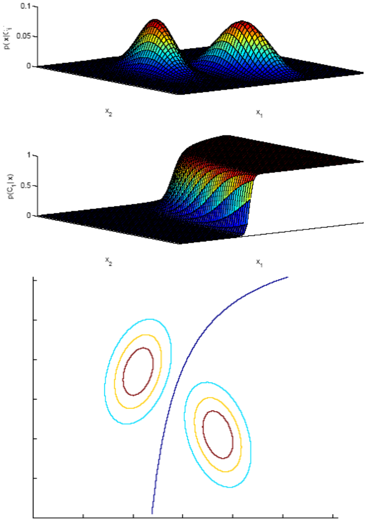
$$
\begin{aligned}
g_i(\mathbf{x}) &= -\frac{1}{2}\log|\mathbf{S}_i| - \frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T\mathbf{S}_i^{-1}(\mathbf{x}-\mathbf{m}_i) + \log\hat{P}(C_i) \\
&= -\frac{1}{2}\log|\mathbf{S}_i| - \frac{1}{2}(\mathbf{x}^T\mathbf{S}_i^{-1}\mathbf{x} - 2\mathbf{x}^T\mathbf{S}_i^{-1}\mathbf{m}_i + \mathbf{m}_i^T\mathbf{S}_i^{-1}\mathbf{m}_i) + \log\hat{P}(C_i) \\
&= \mathbf{x}^T\mathbf{W}_i\mathbf{x} + \mathbf{w}_i^T\mathbf{x} + w_{i0}
\end{aligned}
$$

  where

$$
\begin{aligned}
\mathbf{W}_i &= -\frac{1}{2}\mathbf{S}_i^{-1} \\
\mathbf{w}_i &= \mathbf{S}_i^{-1}\mathbf{m}_i \\
w_{i0} &= -\frac{1}{2}\mathbf{m}_i^T\mathbf{S}_i^{-1}\mathbf{m}_i - \frac{1}{2}\log|\mathbf{S}_i| + \log\hat{P}(C_i)
\end{aligned}
$$

- The discriminant functions are concave and quadratic.
- The decision surface between two categories are hyperquadrics.

Quadratic Discriminant Functions - II



Quadratic Discriminant Functions - III

- The number of parameters to be estimated are Kd for the means and $Kd(d+1)/2$ for the covariance matrices.
- When d is large and samples are small, the estimation is not reliable.
- For the estimates to be reliable on small samples,
  - one may want to decrease dimensionality, d, by redesigning the feature extractor and select a subset of the features or somehow combine existing features.
  - another possibility is to pool the data and estimate a common covariance matrix for all classes.
- If the covariance for different class is different, we call it heteroscedasticity.

Equal Covariance Matrix S - I

- Shared common sample covariance matrix (i.e., homoscedasticity):

$$\mathbf{S} = \sum_i \hat{P}(C_i)\mathbf{S}_i$$

- Discriminant functions are linear:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mathbf{m}_i)^T \mathbf{S}^{-1}(\mathbf{x} - \mathbf{m}_i) + \log \hat{P}(C_i) + \text{const.}$$

- Ignoring terms that are the same for all classes

$$g_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x} + w_{i0}$$

  with

$$\mathbf{w}_i = \mathbf{S}^{-1}\mathbf{m}_i$$

$$w_{i0} = -\frac{1}{2}\mathbf{m}_i^T \mathbf{S}^{-1}\mathbf{m}_i + \log \hat{P}(C_i)$$

- The number of parameters is Kd for the means and $d(d+1)/2$ for the shared covariance matrix.

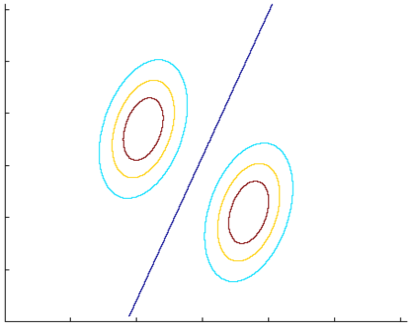
Equal Covariance Matrix S - II

- The decision surfaces for a linear discriminant classifiers are hyperplanes defined by the linear equations $g_i(\mathbf{x}) = g_j(\mathbf{x})$.
  - The equation can be written as

$$
\begin{aligned}
&(\mathbf{w}_i - \mathbf{w}_j)^T \mathbf{x} + w_{i0} - w_{j0} = 0 \\
&\mathbf{w}^T (\mathbf{x} - \mathbf{x}_0) = 0 \\
&\mathbf{w} = \mathbf{S}^{-1}(\mathbf{m}_i - \mathbf{m}_j) \\
&\mathbf{x}_0 = \frac{1}{2}\mathbf{S}^{-1}(\mathbf{m}_i + \mathbf{m}_j) - \frac{1}{\|\mathbf{S}^{-1}(\mathbf{m}_i - \mathbf{m}_j)\|^2} \log \frac{\hat{P}(C_i)}{\hat{P}(C_j)} \mathbf{S}^{-1}(\mathbf{m}_i - \mathbf{m}_j)
\end{aligned}
$$

  - These equations define a hyperplane through point $\mathbf{x}_0$ with a normal vector $\mathbf{w}$.
- If the priors are equal, the optimal decision rule is to assign input to the class whose mean's Mahalanobis distance to the input is the smallest.
- Unequal priors shift the boundary toward the less likely class.

Equal Covariance Matrix S - III



Decision regions of such a linear classifier are convex.

Equal and Diagonal S - I

- Naive Bayes' classifier: if the variables are independent, $\mathbf{\Sigma}$ becomes a diagonal matrix.
- Class-conditional densities:

$$p(\mathbf{x} \mid C_i) = \prod_j p(x_j \mid C_i)$$

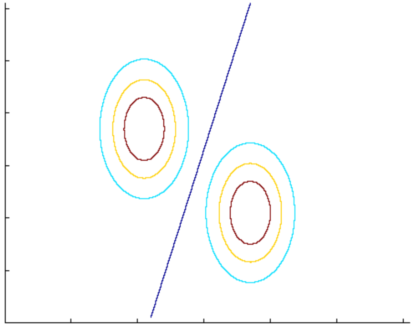
where $p(x_j \mid C_i)$ are univariate Gaussian distributions.

- Discriminant functions:

$$g_i(\mathbf{x}) = -\frac{1}{2}\sum_{j=1}^{d}\left(\frac{x_j - m_{ij}}{s_j}\right)^2 + \log \hat{P}(C_i)$$

- Classification based on weighted Euclidean distance.
- The number of parameters is Kd for the means and d for the variances.

Equal and Diagonal S - II



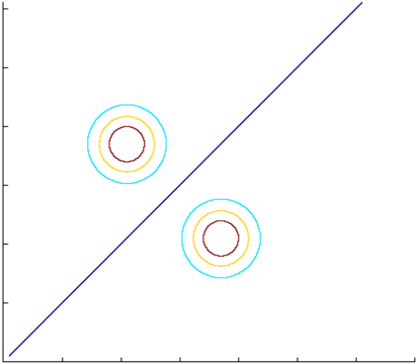
Equal and Diagonal S with Equal Variances - I

- If we assume further that all variances are equal, i.e., $\boldsymbol{\Sigma} = s^2\mathbf{I}$, weighted Euclidean distance reduces to Euclidean distance.
- Discriminant functions:

$$g_i(\mathbf{x}) = -\frac{1}{2s^2}\|\mathbf{x} - \mathbf{m}_i\|^2 + \log \hat{P}(C_i)$$
$$= -\frac{1}{2s^2}\sum_{j=1}^{d}(x_j - m_{ij})^2 + \log \hat{P}(C_i)$$

- Discriminant functions are linear.
- The number of parameters in this case is Kd for the means and 1 for s^2.
- If the priors are equal, we have $g_i(\mathbf{x}) = -\|\mathbf{x} - \mathbf{m}_i\|^2$
  - nearest mean classifier: it assigns the input to the class of the nearest mean
  - template matching procedure: each mean acts as a prototype/template for the class.

Equal and Diagonal S with Equal Variances - II

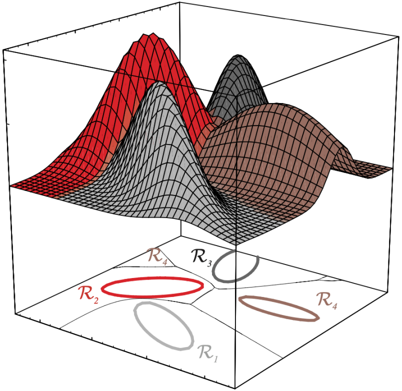


Tuning Model Complexity

Assumption	Covariance matrix	No. of parameters
Shared, hyperspherical	$\mathbf{S}_i = \mathbf{S} = s^2\mathbf{I}$	1
Shared, axis-aligned	$\mathbf{S}_i = \mathbf{S}$, with $s_{ij} = 0$	d
Shared, hyperellipsoidal	$\mathbf{S}_i = \mathbf{S}$	$d(d+1)/2$
Different, hyperellipsoidal	$\mathbf{S}_i$	$Kd(d+1)/2$

- Complexity increases (i.e., less restricted $\mathbf{S}$)
  $\Rightarrow$ bias decreases and variance increases
- Regularization: uses strong bias to control model complexity.

General Case for Multiple Classes



Outline

Discrete Features: Bernoulli

- Bernoulli (or binary) variables x_j:

$$p_{ij} \equiv p(x_j = 1 \mid C_i)$$

- If x_j's are independent given C_i (i.e, naive Bayes'):

$$p(\mathbf{x} \mid C_i) = \prod_{j=1}^{d} p_{ij}^{x_j} (1 - p_{ij})^{1-x_j}$$

  giving linear discriminant functions:

$$\begin{aligned} g_i(\mathbf{x}) &= \log p(\mathbf{x} \mid C_i) + \log P(C_i) \\ &= \sum_j \Big[x_j \log p_{ij} + (1 - x_j) \log(1 - p_{ij}) \Big] + \log P(C_i) \end{aligned}$$

- Given sample $\mathcal{X} = \{\mathbf{x}^t\}_{t=1}^{N}$, the maximum likelihood estimators:

$$\hat{p}_{ij} = \frac{\sum_t x_j^t r_i^t}{\sum_t r_i^t}$$

Discrete Features: Generalized Bernoulli

- Generalized Bernoulli (or multinomial) variables $x_j \in \{v_1, \ldots, v_{n_j}\}$
- Indicator variables:

$$z_{jk} = \begin{cases} 1 & \text{if } x_j = v_k \\ 0 & \text{otherwise} \end{cases}$$

- Define

$$p_{ijk} \equiv p(z_{jk} = 1 \mid C_i) = p(x_j = v_k \mid C_i)$$

- If x_j's are independent:

$$p(\mathbf{x} \mid C_i) = \prod_{j=1}^{d} \prod_{k=1}^{n_j} p_{ijk}^{z_{jk}}$$

$$g_i(\mathbf{x}) = \sum_j \sum_k z_{jk} \log p_{ijk} + \log P(C_i)$$

- Given sample $\mathcal{X} = \{\mathbf{x}^t\}_{t=1}^N$, the maximum likelihood estimators:

$$\hat{p}_{ijk} = \frac{\sum_t z_{jk}^t r_i^t}{\sum_t r_i^t}$$

Outline

Multivariate Linear Regression

- Multivariate linear regression:
$$r = f(\mathbf{x}) + \epsilon$$
where $f(\mathbf{x}) \approx$ estimator $g(\mathbf{x} \mid w_0, w_1, \ldots, w_d) = w_0 + w_1 x_1 + \cdots + w_d x_d$.
- In some literature (especially statistical literature), this is called multiple linear regression; statisticians use the term multivariate when there are multiple outputs.
- Given $\mathcal{X} = \{(\mathbf{x}^t, r^t)\}_{t=1}^N$, error function:
$$E(w_0, w_1, \ldots, w_d \mid \mathcal{X}) = \frac{1}{2} \sum_t \left(r^t - w_0 - w_1 x_1^t - \cdots - w_d x_d^t \right)^2$$
- Maximizing the Gaussian likelihood is equivalent to minimizing the sum of squared errors.

Normal Equations

- Taking the derivative with respect to the parameters, we get the normal equations for multivariate linear regression:

$$\mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{r}$$

where

$$\mathbf{X} = \begin{bmatrix} 1 & x_1^1 & x_2^1 & \cdots & x_d^1 \\ 1 & x_1^2 & x_2^2 & \cdots & x_d^2 \\ \vdots & & & & \\ 1 & x_1^N & x_2^N & \cdots & x_d^N \end{bmatrix}$$

$$\mathbf{w} = (w_0, w_1, \ldots, w_d)^T$$

$$\mathbf{r} = (r^1, r^2, \ldots, r^N)^T$$

- Estimated parameters (assuming that $\mathbf{X}^T\mathbf{X}$ is invertible):

$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{r}$$

Multivariate Polynomial Regression

- Define new higher-order variables, e.g.

$$z_1 = x_1,\; z_2 = x_2\; z_3 = (x_1)^2,\; z_4 = (x_2)^2,\; z_5 = x_1 x_2$$

- Apply multivariate linear regression in the new $\mathbf{z}$ space.
- Actually using higher-order terms of inputs as additional inputs is only one possibility; we can define any nonlinear function of the original inputs using basis functions, like $z = \sin(x)$.
- This idea of generalizing the linear model is frequently used in later course.