Machine Learning

Math Essentials
Part 2

Gaussian distribution

- Most commonly used continuous probability distribution

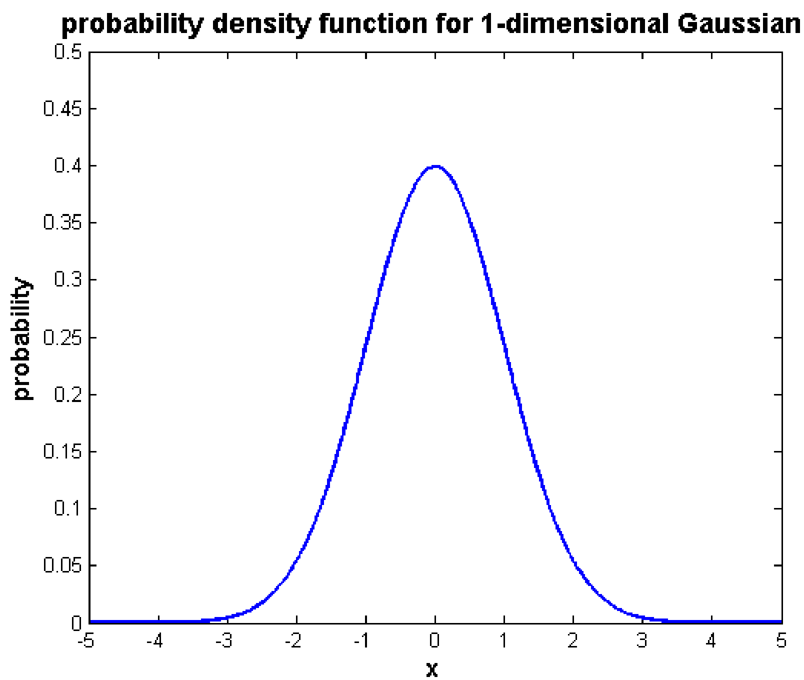
- Also known as the normal distribution

- Two parameters define a Gaussian:
  - Mean μ location of center
  - Variance σ^2 width of curve

Gaussian distribution

In one dimension

$$N(x \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$



Gaussian distribution

In one dimension

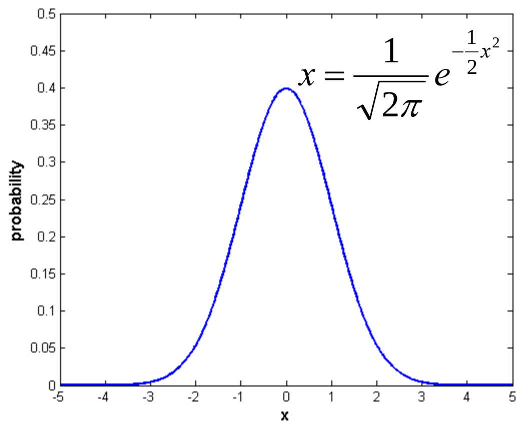
Causes pdf to decrease as distance from center increases

$$N(x \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

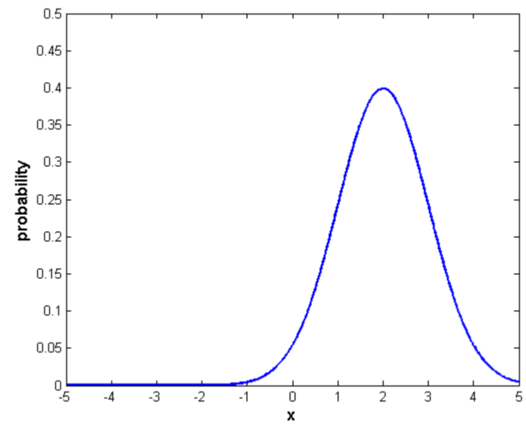
Controls width of curve

Normalizing constant: insures that distribution integrates to 1

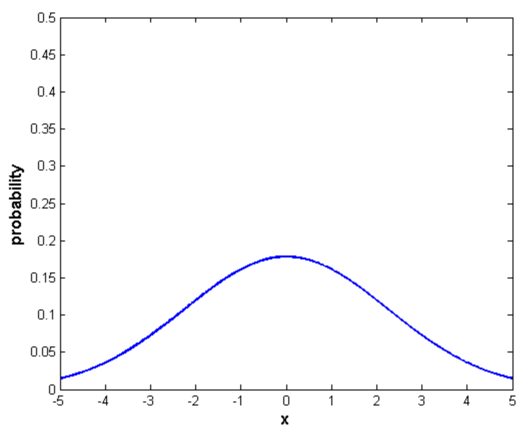
Gaussian distribution



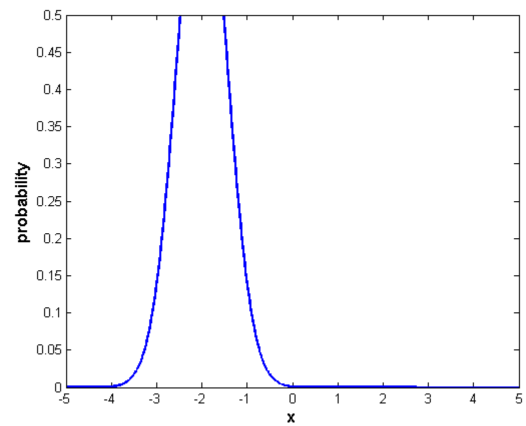
μ = 0   σ² = 1



μ = 2   σ² = 1



μ = 0   σ² = 5



μ = -2   σ² = 0.3

Multivariate Gaussian distribution

In *d* dimensions

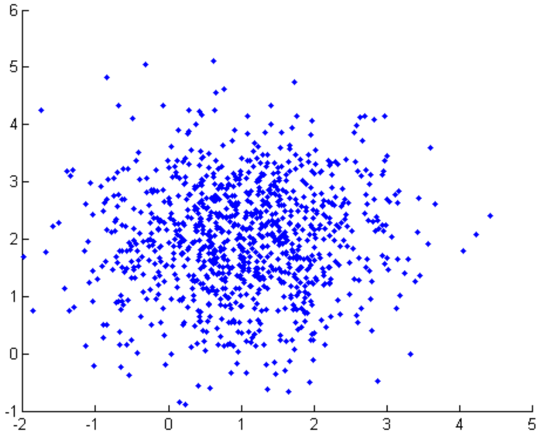
$$N(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$

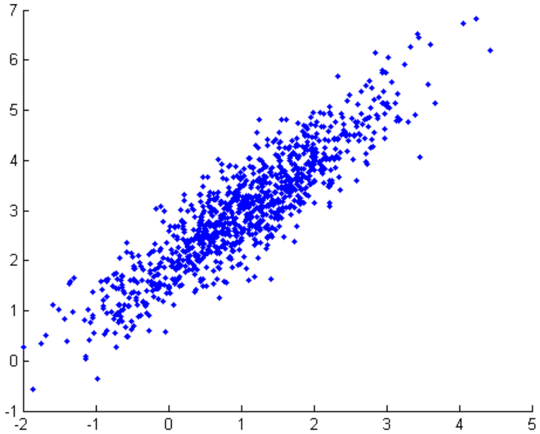
- **x** and μ now *d*-dimensional vectors
  - μ gives center of distribution in *d*-dimensional space
- σ^2 replaced by Σ, the *d* x *d* covariance matrix
  - Σ contains pairwise covariances of every pair of features
  - Diagonal elements of Σ are variances σ^2 of individual features
  - Σ describes distribution's shape and spread

Multivariate Gaussian distribution

- Covariance
  - Measures tendency for two variables to deviate from their means in same (or opposite) directions at same time

no covariance

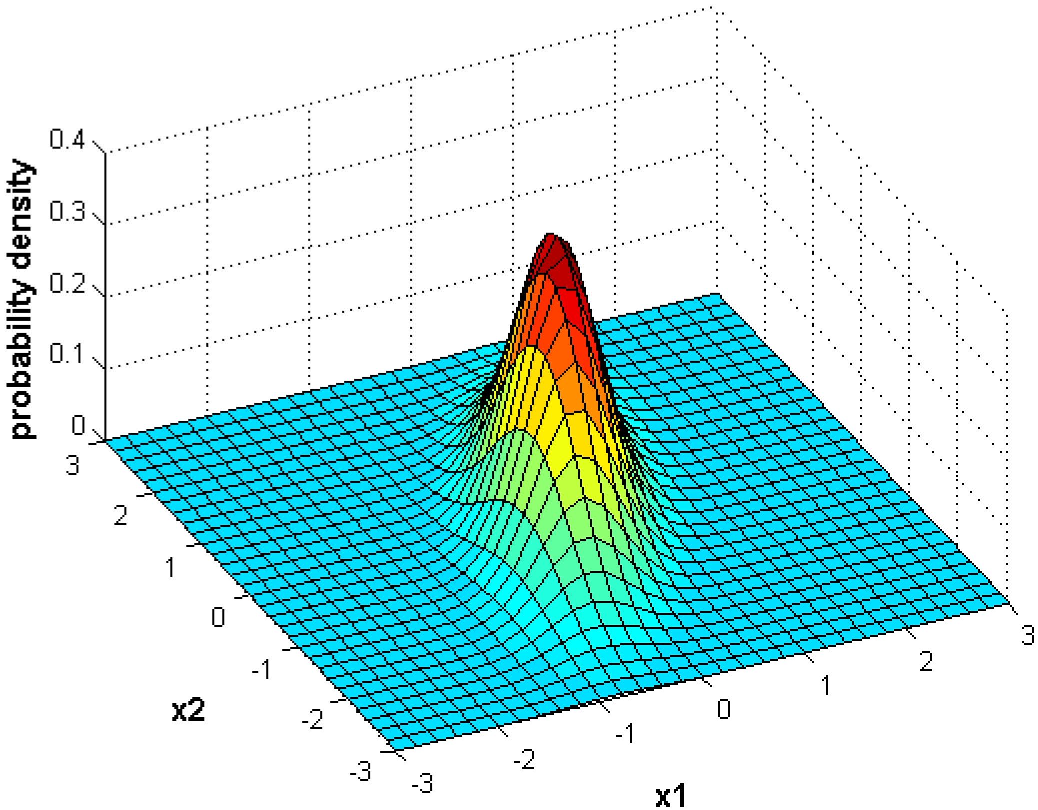




high (positive) covariance

Multivariate Gaussian distribution

In two dimensions

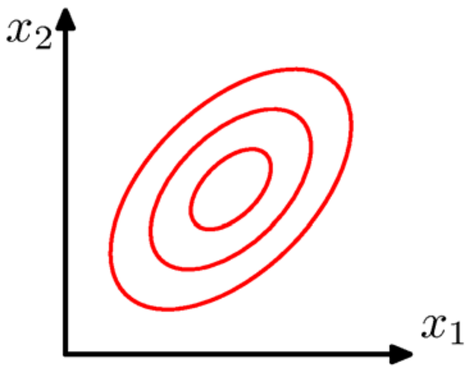


$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

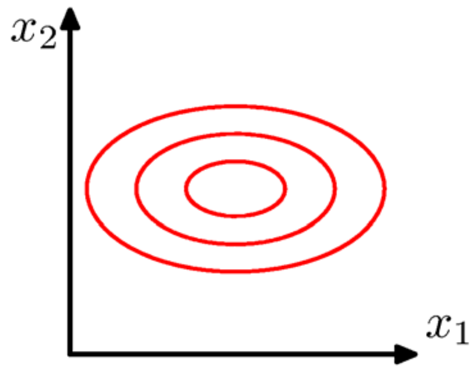
$$\boldsymbol{\Sigma} = \begin{bmatrix} 0.25 & 0.3 \\ 0.3 & 1 \end{bmatrix}$$

Multivariate Gaussian distribution

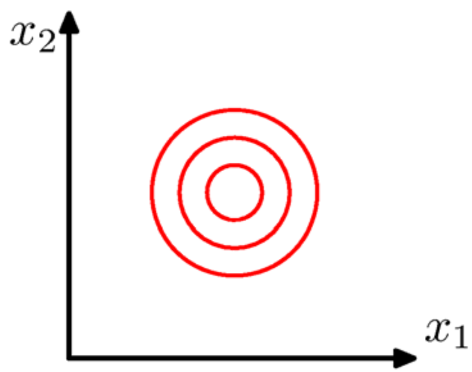
In two dimensions



(a)



(b)



(c)

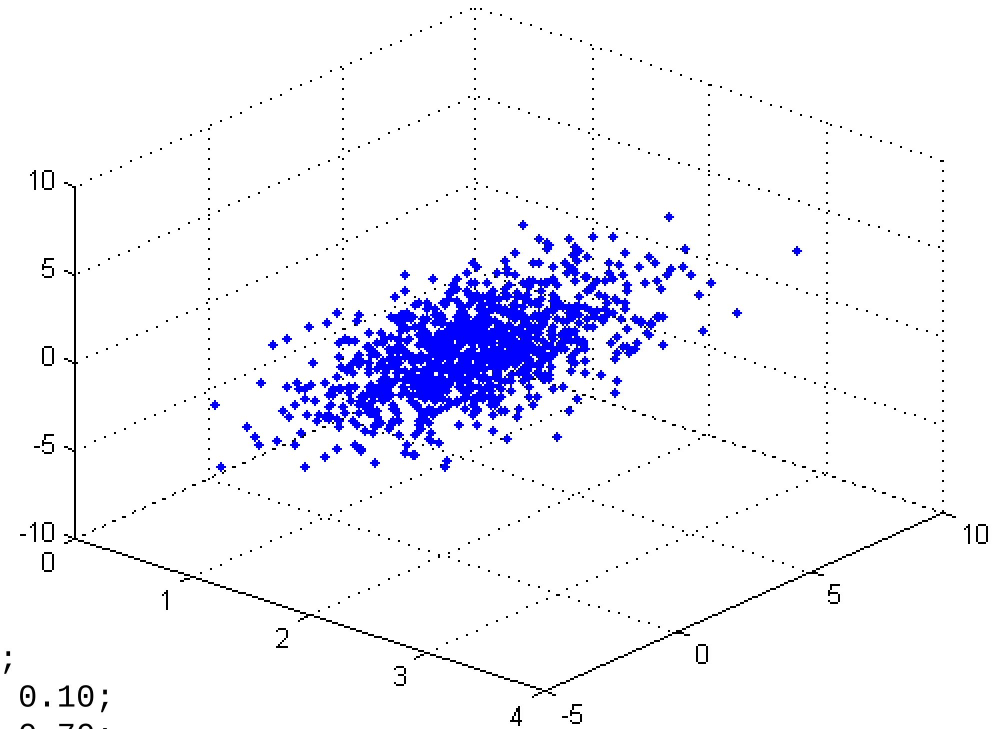
$$\mathbf{\Sigma} = \begin{bmatrix} 2 & 0.6 \\ 0.6 & 2 \end{bmatrix}$$

$$\mathbf{\Sigma} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Multivariate Gaussian distribution

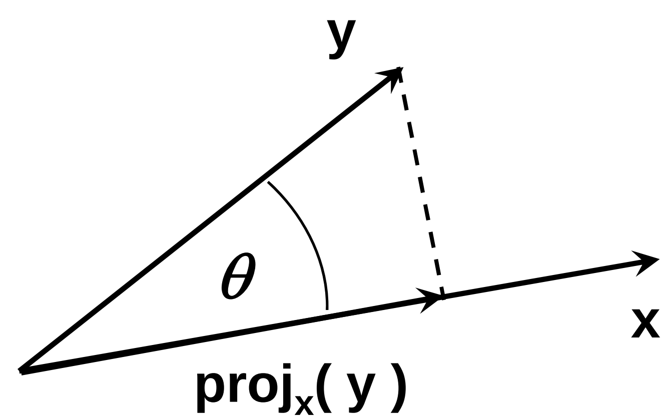
In three dimensions




```
rng( 1 );
mu =    [ 2; 1; 1 ];
sigma = [ 0.25 0.30 0.10;
          0.30 1.00 0.70;
          0.10 0.70 2.00] ;
x = randn( 1000, 3 );
x = x * sigma;
x = x + repmat( mu', 1000, 1 );
scatter3( x( :, 1 ), x( :, 2 ), x( :, 3 ), '.' );
```

Vector projection

- Orthogonal projection of **y** onto **x**
  - Can take place in any space of dimensionality ≥ 2
  - Unit vector in direction of **x** is
    $\mathbf{x} \,/\, \|\, \mathbf{x} \,\|$
  - Length of projection of **y** in direction of **x** is
    $\|\, \mathbf{y} \,\| \cdot \cos(\theta)$
  - Orthogonal projection of **y** onto **x** is the vector



$$\mathbf{proj}_{\mathbf{x}}(\,\mathbf{y}\,) \;=\; \mathbf{x} \cdot \|\,\mathbf{y}\,\| \cdot \cos(\theta) \,/\, \|\,\mathbf{x}\,\| \;=$$
$$[\,(\,\mathbf{x} \cdot \mathbf{y}\,) \,/\, \|\,\mathbf{x}\,\|^2\,]\,\mathbf{x}$$ (using dot product alternate form)

Linear models

- There are many types of linear models in machine learning.
  - Common in both classification and regression.
  - A linear model consists of a vector **w** in *d*-dimensional feature space.
  - The vector **w** attempts to capture the strongest gradient (rate of change) in the output variable, as seen across all training samples.
  - Different linear models optimize **w** in different ways.
  - A point **x** in feature space is mapped from *d* dimensions to a scalar (1-dimensional) output *z* by projection onto **w:**

$$z = w_0 + \mathbf{w} \cdot \mathbf{x} = w_0 + w_1 x_1 + \cdots + w_d x_d$$

cf. Lecture 5b
$w_0 \equiv \alpha$
$\mathbf{w} \equiv \beta$

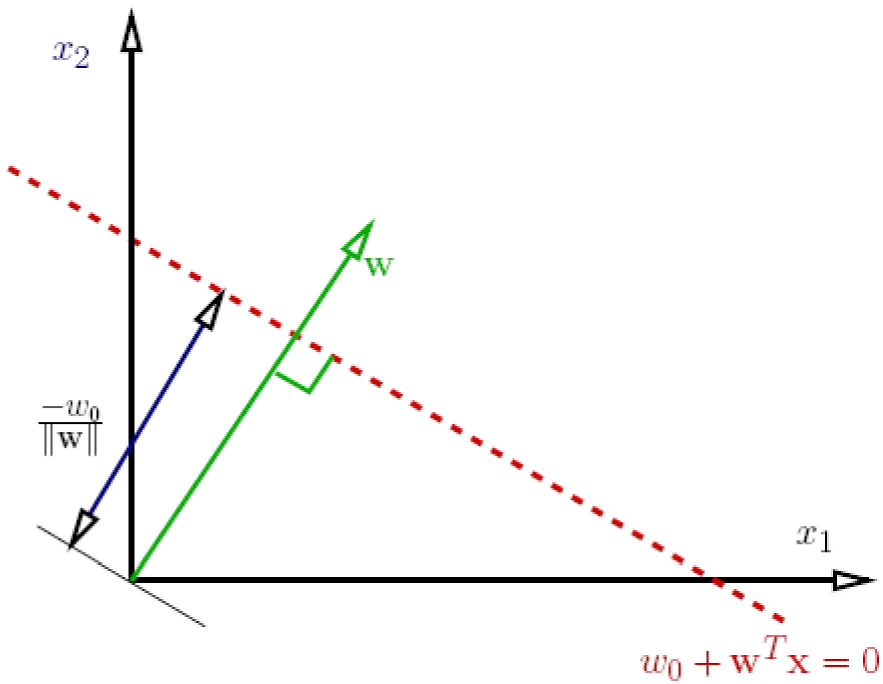
Linear models

- There are many types of linear models in machine learning.
  - The projection output z is typically transformed to a final predicted output y by some function f:

$$y = f(z) = f(w_0 + \mathbf{w} \cdot \mathbf{x}) = f(w_0 + w_1 x_1 + \cdots + w_d x_d)$$

    - example: for logistic regression, f is logistic function
    - example: for linear regression, $f(z) = z$
  - Models are called linear because they are a linear function of the model vector components $w_1, \ldots, w_d$.
  - Key feature of all linear models: no matter what f is, a constant value of z is transformed to a constant value of y, so decision boundaries remain linear even after transform.

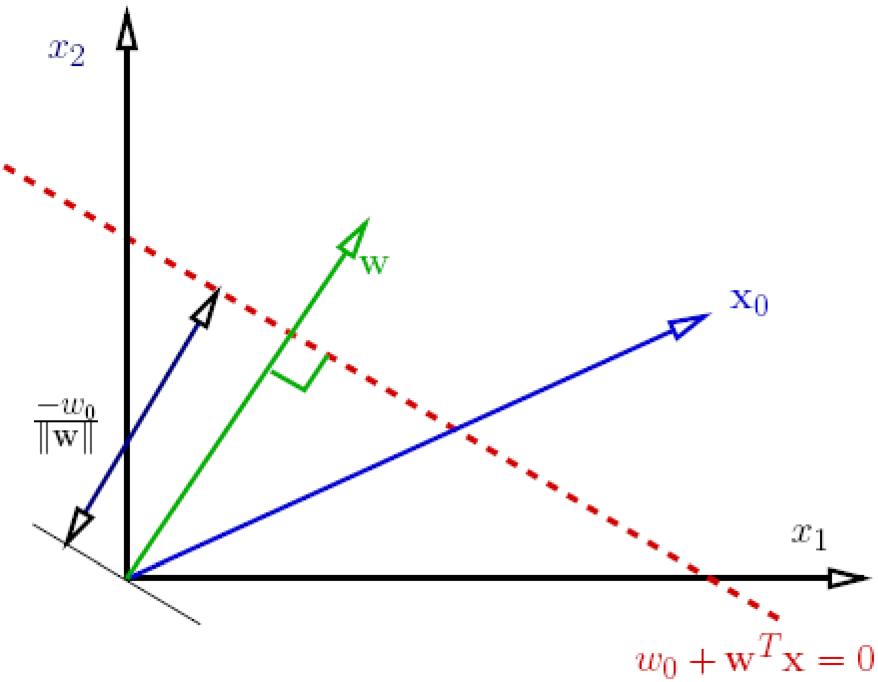
Geometry of projections



- $\mathbf{w}^T\mathbf{x} = 0$: a line passing through the origin and *orthogonal* to $\mathbf{w}$
- $\mathbf{w}^T\mathbf{x} + w_0 = 0$ shifts the line along $\mathbf{w}$.

slide thanks to Greg Shakhnarovich (CS195-5, Brown Univ., 2006)

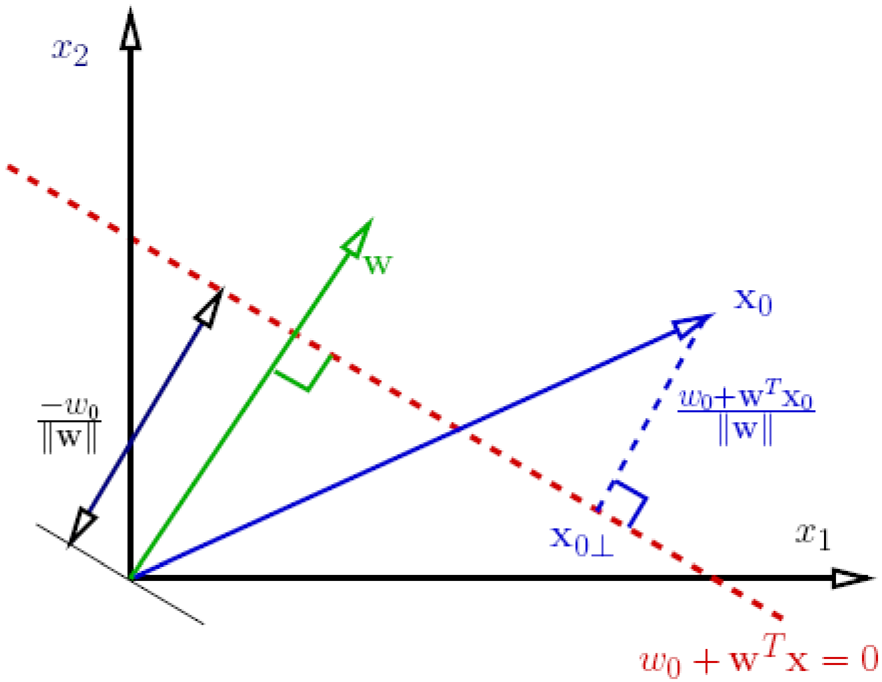
Geometry of projections



- $\mathbf{w}^T\mathbf{x} = 0$: a line passing through the origin and *orthogonal* to $\mathbf{w}$
- $\mathbf{w}^T\mathbf{x} + w_0 = 0$ shifts the line along $\mathbf{w}$.

slide thanks to Greg Shakhnarovich (CS195-5, Brown Univ., 2006)

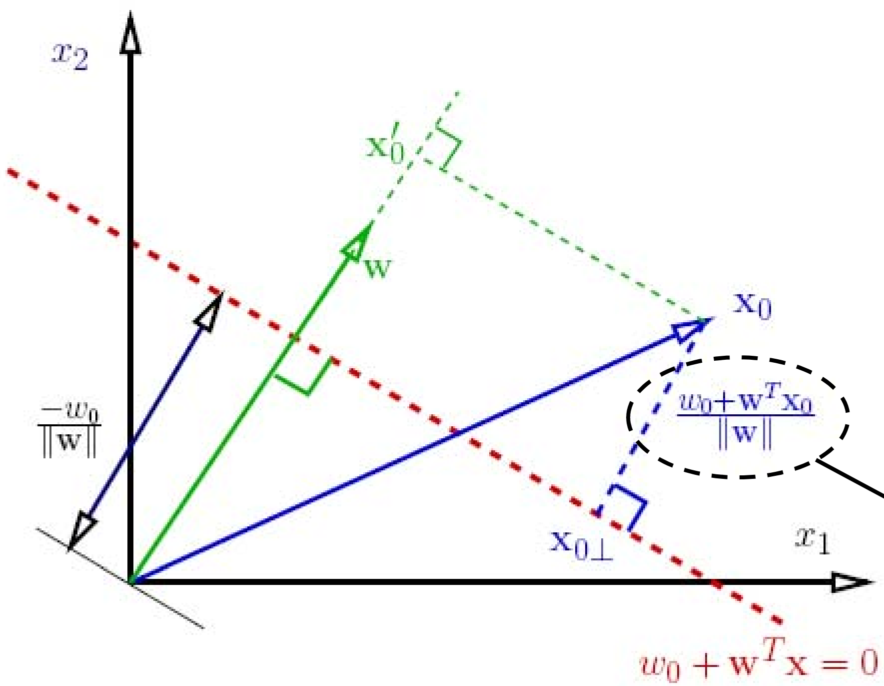
Geometry of projections



- $\mathbf{w}^T\mathbf{x} = 0$: a line passing through the origin and *orthogonal* to $\mathbf{w}$
- $\mathbf{w}^T\mathbf{x} + w_0 = 0$ shifts the line along $\mathbf{w}$.

slide thanks to Greg Shakhnarovich (CS195-5, Brown Univ., 2006)

Geometry of projections



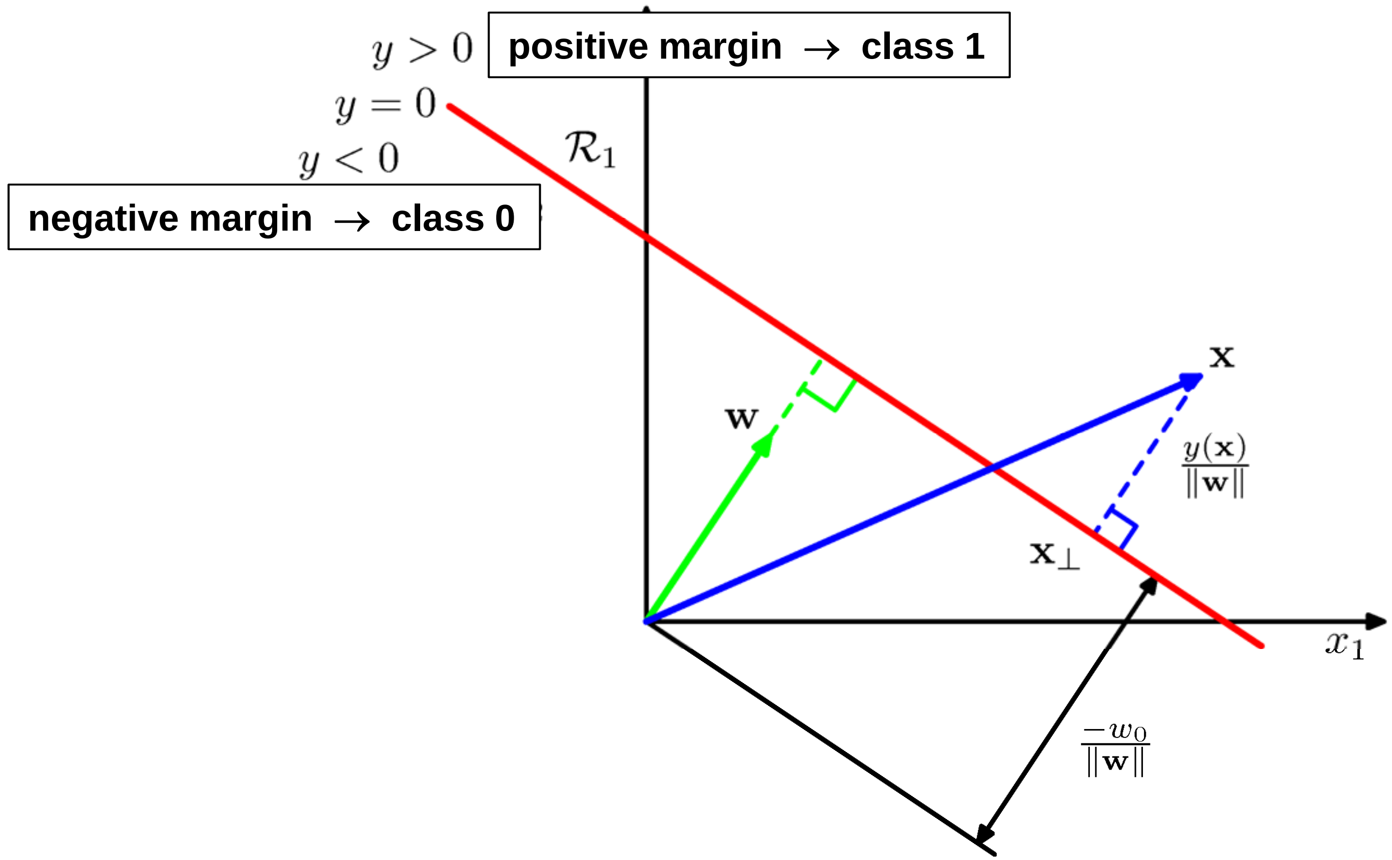
- $\mathbf{w}^T\mathbf{x} = 0$: a line passing through the origin and *orthogonal* to $\mathbf{w}$
- $\mathbf{w}^T\mathbf{x} + w_0 = 0$ shifts the line along $\mathbf{w}$.

margin

- $\mathbf{x}'$ is the projection of $\mathbf{x}$ on $\mathbf{w}$.
- Set up a new 1D coordinate system: $\mathbf{x} \to (w_0 + \mathbf{x}^T\mathbf{x})/\|\mathbf{w}\|$.

slide thanks to Greg Shakhnarovich (CS195-5, Brown Univ., 2006)

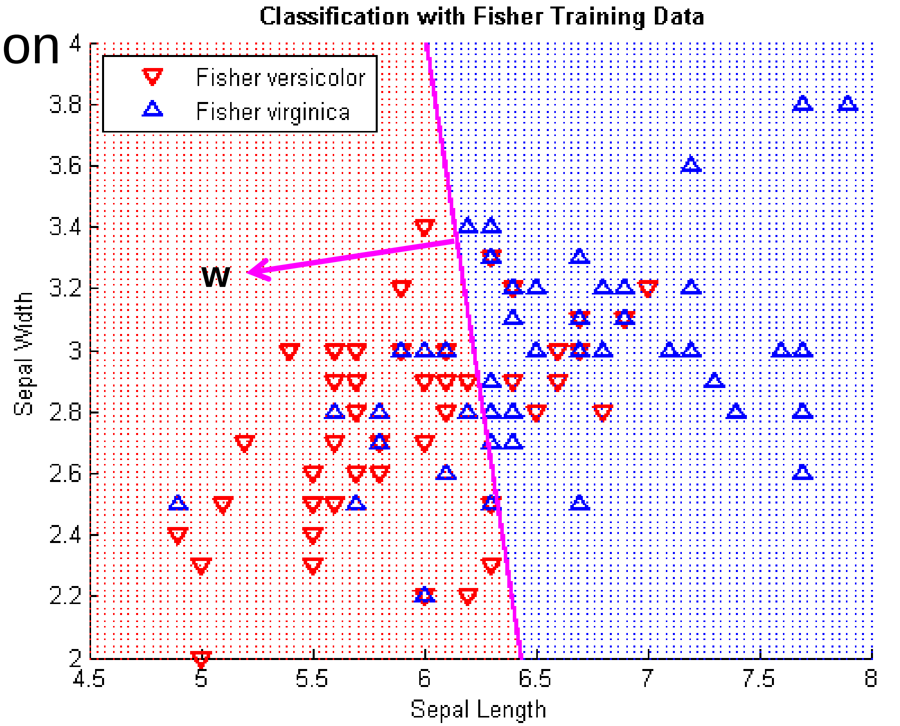
From projection to prediction



Logistic regression in two dimensions

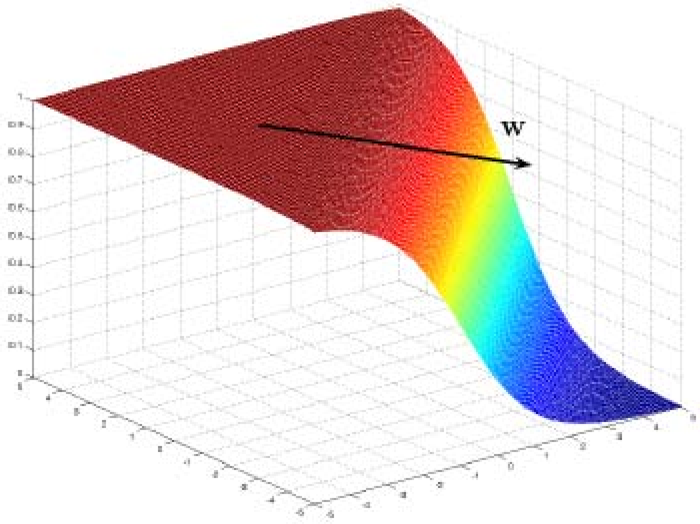
Interpreting the model vector of coefficients

- From MATLAB: `B = [ 13.0460   -1.9024   -0.4047 ]`
- w_0 = B(1), $\mathbf{w} = [\, w_1\ w_2 \,]$ = B(2 : 3)
- w_0, **w** define location and orientation of decision boundary
  – $-w_0$ is distance of decision boundary from origin
  – decision boundary is perpendicular to **w**
- magnitude of **w** defines gradient of probabilities between 0 and 1



Logistic function in *d* dimensions

- What if $\mathbf{x} \in \mathbb{R}^d \; = [x_1 \ldots \; x_d]^T$?
- $\sigma(w_0 + \mathbf{w}^T\mathbf{x})$ is a scalar function of a scalar variable $w_0 + \mathbf{w}^T\mathbf{x}$.



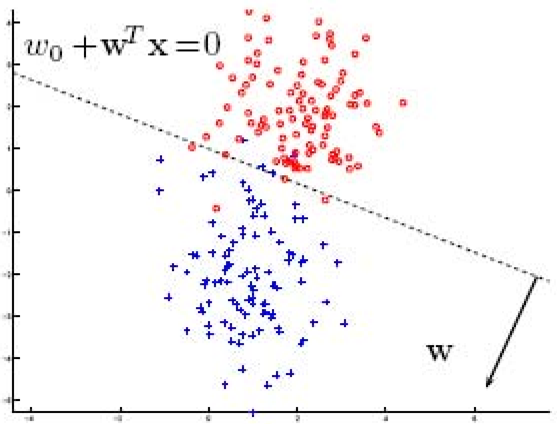
- the direction of $\mathbf{w}$ determines orientation;
- w_0 determines the location;
- $\|\mathbf{w}\|$ determines the slope.

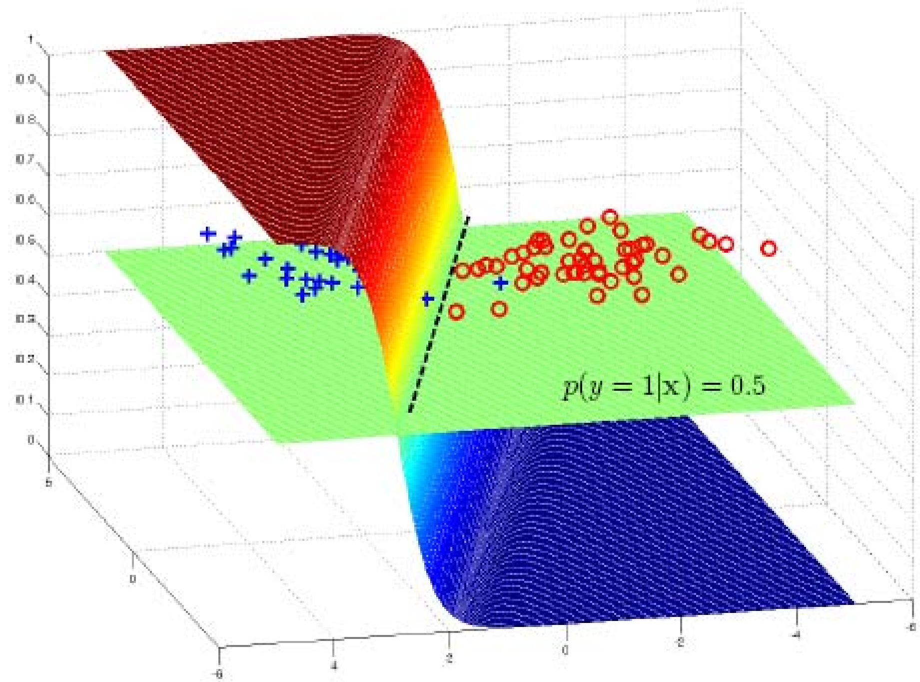
slide thanks to Greg Shakhnarovich (CS195-5, Brown Univ., 2006)

Decision boundary for logistic regression

$$p\,(y=1\,|\,\mathbf{x}) \;=\; \sigma(w_0+\mathbf{w}^T\mathbf{x}) = 1/2 \;\Leftrightarrow\; w_0+\mathbf{w}^T\mathbf{x}=0$$

- With linear logistic model we get a linear decision boundary.





slide thanks to Greg Shakhnarovich (CS195-5, Brown Univ., 2006)