# Machine learning - HT 2016
# 3. Maximum Likelihood

Varun Kanade

University of Oxford
January 27, 2016

# Outline

## Probabilistic Framework

- Formulate linear regression in the language of probability
- Introduce the maximum likelihood estimate
- Relation to least squares estimate

## Basics of Probability

- Univariate and multivariate normal distribution
- Laplace distribution
- Likelihood, Entropy and its relation to learning

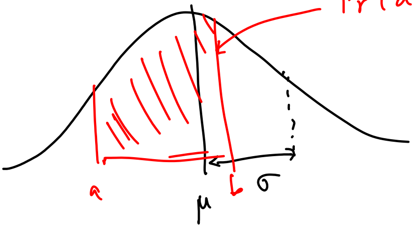
# Univariate Gaussian (Normal) Distribution

The univariate normal distribution is defined by the following density function

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad\qquad X \sim \mathcal{N}(\mu, \sigma^2)$$

Here $\mu$ is the mean and $\sigma^2$ is the variance.

$Pr(a < x < b)$




Recall:

$$\int_{-\infty}^{+\infty} p(x)\,dx = 1$$

$$\int_{-\infty}^{+\infty} x\,p(x)\,dx = \mu$$

$$\int_{-\infty}^{+\infty} x^2 p(x)\,dx - \mu^2 = \sigma^2$$

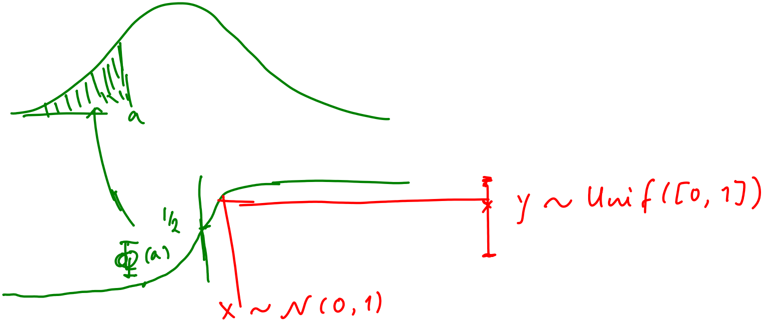
# Sampling from a Gaussian distribution

Sampling from $X \sim \mathcal{N}(\mu, \sigma^2)$

By setting $Y = \frac{X-\mu}{\sigma}$, sample from $Y \sim \mathcal{N}(0,1)$

Cumulative distribution function

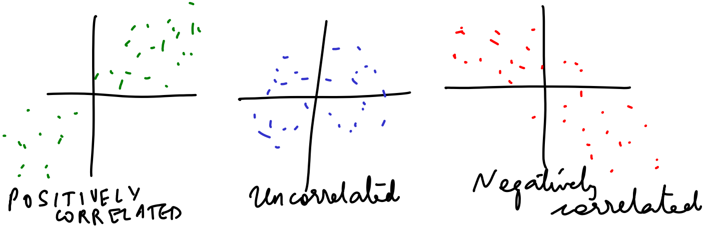
$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$



# Covariance and Correlation

For random variable $X$ and $Y$ the covariance measures how the random variable change jointly.

$$\text{cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$




Covariance depends on the scale of the random variable. The (Pearson) correlation coefficient normalizes the covariance to give a value between $-1$ and $+1$.

$$\text{corr}(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y},$$

where $\sigma_X^2 = \mathbb{E}[(X - \mathbb{E}[X])^2]$ and $\sigma_Y^2 = \mathbb{E}[(Y - \mathbb{E}[Y])^2]$.

# Multivariate Gaussian Distribution

Suppose $\mathbf{x}$ is a $n$-dimensional random vector. The covariance matrix consists of all pariwise covariances.

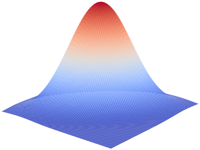
$$\operatorname{cov}(\mathbf{x}) = \mathbb{E}\left[(\mathbf{x}-\mathbb{E}[\mathbf{x}])(\mathbf{x}-\mathbb{E}[\mathbf{x}])^T\right] = \begin{bmatrix} \operatorname{var}(X_1) & \operatorname{cov}(X_1, X_2) & \cdots & \operatorname{cov}(X_1, X_n) \\ \operatorname{cov}(X_2, X_1) & \operatorname{var}(X_2) & \cdots & \operatorname{cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(X_n, X_1) & \operatorname{cov}(X_n, X_2) & \cdots & \operatorname{var}(X_n, X_n) \end{bmatrix}.$$

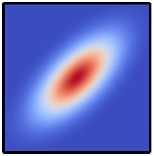
If $\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}]$ and $\boldsymbol{\Sigma} = \operatorname{cov}[\mathbf{x}]$, the multivariate normal is defined by the density

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

$\Sigma$ is positive definite if $\forall y \neq 0 \in \mathbb{R}^n$,

$y^T \Sigma y > 0$





# Bivariate Gaussian Distribution

Suppose $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$

What is the joint probability distribution $p(x_1, x_2)$?

$$p(x_1, x_2) = p(x_1) \cdot p(x_2) \quad \text{(if } x_1 \text{ \& } x_2 \text{ independent)}$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}} \cdot \frac{1}{\sqrt{2\pi}\,\sigma_2} \cdot e^{-\frac{(x_2-\mu_2)^2}{2\sigma_2^2}}$$

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \qquad = \frac{1}{(2\pi)\,|\Sigma|^{1/2}} \exp\left(-\frac{1}{2} [x_1-\mu_1, x_2-\mu_2]\, \Sigma^{-1} \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix}\right)$$

$\boxed{\text{general case, by rotating (axes)}}$

AXIS-ALIGNED

The locus of points having same prob density are ellipses around $\mu$.

$$\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} = \underline{\text{Constant}}$$

Suppose you are given three <u>independent</u> samples:
$x_1 = 1, \ \ x_2 = 2.7, \ \ x_3 = 3.$

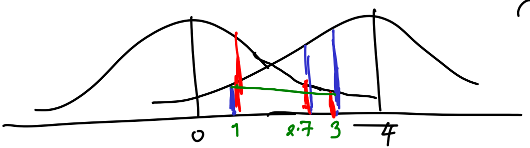
You know that the data were generated from $\mathcal{N}(0, 1)$ or $\mathcal{N}(4, 1)$.

Let $\boldsymbol{\theta}$ represent the parameters of the distribution. Then the probability of observing data with parameter $\boldsymbol{\theta}$ is called the likelihood:

$$p(x_1, x_2, x_3 \mid \boldsymbol{\theta}) = p(x_1 \mid \boldsymbol{\theta}) p(x_2 \mid \boldsymbol{\theta}) p(x_3 \mid \boldsymbol{\theta})$$

We have to chose between $\boldsymbol{\theta} = 0$ and $\boldsymbol{\theta} = 4$. Which one?

$\mathcal{N}(0,4)$ is more likely




Maximum Likelihood Estimation (MLE): Pick $\boldsymbol{\theta}$ that maximizes the likelihood.

# Linear Regression

Recall our linear regression model

$$y = \mathbf{x}^T \mathbf{w} + \text{noise}$$

Model $y$ (conditioned on $\mathbf{x}$) as a random variable. Given $\mathbf{x}$ and the model parameter $\mathbf{w}$:

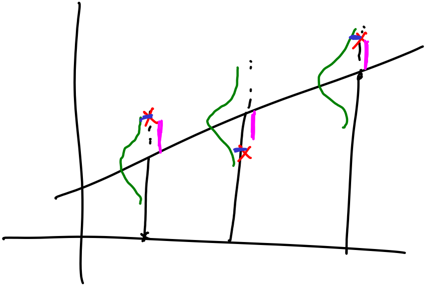
$$\mathbb{E}[y \mid \mathbf{x}, \mathbf{w}] = \mathbf{x}^T \mathbf{w}$$

We can be more specific in choosing our model for $y$. Let us assume that given $\mathbf{x}, \mathbf{w}$, $y$ is Gaussian with mean $\mathbf{x}^T \mathbf{w}$ and variance $\sigma^2$.

$$y \sim \mathcal{N}(\mathbf{x}^T \mathbf{w}, \sigma^2) = \mathbf{x}^T \mathbf{w} + \mathcal{N}(0, \sigma^2)$$

# Likelihood of Linear Regression

Suppose we observe data $\langle (\mathbf{x}_i, y_i) \rangle_{i=1}^{m}$. What is the likelihood of observing the data for model parameters $\mathbf{w}$, $\sigma$?



Likelihood = product of – lines

Least Square:

Sum of squares of lengths of | lines

# Likelihood of Linear Regression

Suppose we observe data $\langle(\mathbf{x}_i, y_i)\rangle_{i=1}^m$. What is the likelihood of observing the data for model parameters $\mathbf{w}$, $\sigma$?

$$p\left(y_1, \ldots, y_m \mid \mathbf{x}_1, \ldots, \mathbf{x}_m, \mathbf{w}, \sigma\right) = \prod_{i=1}^m p\left(y_i \mid \mathbf{x}_i, \mathbf{w}, \sigma\right)$$

Recall that $y_i \sim \mathbf{x}_i^T \mathbf{w} + \mathcal{N}(0, \sigma^2)$. So

$$\begin{aligned} p\left(y_1, \ldots, y_m \mid \mathbf{x}_1, \ldots, \mathbf{x}_m, \mathbf{w}, \sigma\right) &= \prod_{i=1}^m \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \mathbf{x}_i^T \mathbf{w})^2}{2\sigma^2}} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{m/2} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^m (y_i - \mathbf{x}_i^T \mathbf{w})^2} \end{aligned}$$

Want to find parameters $\mathbf{w}$ and $\sigma$ that maximize the likelihood

# Likelihood of Linear Regression

It is simpler to look at the log-likelihood. Taking logs

$$\mathrm{LL}(y_1, \ldots, y_m \mid \mathbf{x}_1, \ldots, \mathbf{x}_m, \mathbf{w}, \sigma) = -\frac{m}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{m} (\mathbf{x}_i^T \mathbf{w} - y_i)^2$$

$$\mathrm{LL}(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \sigma) = -\frac{m}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (\mathbf{X}\mathbf{w} - \mathbf{y})^T (\mathbf{X}\mathbf{w} - \mathbf{y})$$

How to find $\mathbf{w}$ that maximizes the likelihood?

$$\nabla_w LL = -\frac{1}{2\sigma^2}\left(2X^TXw - 2X^Ty\right) = 0$$

$$w = (X^TX)^{-1}X^Ty$$ (same as least square estimate)

Make sure that Hessian is negative definite to be sure that you have a maxima

# Maximum Likelihood and Least Squares

Let us in fact look at negative log-likelihood (which is more like loss)

$$\mathrm{NLL}(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \sigma) = \frac{1}{2\sigma^2}(\mathbf{X}\mathbf{w} - \mathbf{y})^T(\mathbf{X}\mathbf{w} - \mathbf{y}) + \frac{m}{2}\log(2\pi\sigma^2)$$

And recall the squared loss objective

$$L(\mathbf{w}) = (\mathbf{X}\mathbf{w} - \mathbf{y})^T(\mathbf{X}\mathbf{w} - \mathbf{y})$$

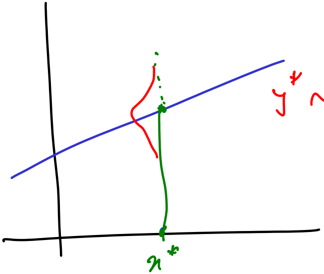
We can also find the MLE for $\sigma$. As exercise show that the MLE of $\sigma$ is

$$\sigma^2_{\mathsf{ML}} = \frac{1}{m}(\mathbf{X}\mathbf{w}_{\mathsf{ML}} - \mathbf{y})^T(\mathbf{X}\mathbf{w}_{\mathsf{ML}} - \mathbf{y})$$

# Making Prediction

Given training data $\mathcal{D} = \langle(\mathbf{x}_i, y_i)\rangle_{i=1}^m$ we can use MLE estimators to make predictions on new points and also give confidence intervals.

$$y_{\text{new}} \mid \mathbf{x}_{\text{new}}, \mathcal{D} \sim \mathcal{N}(\mathbf{x}_{\text{new}}^T \mathbf{w}_{\text{ML}}, \sigma_{\text{ML}}^2)$$




$y^* \sim \mathcal{N}((x^*_{new})^T \hat{w}, \hat{\sigma}^2)$

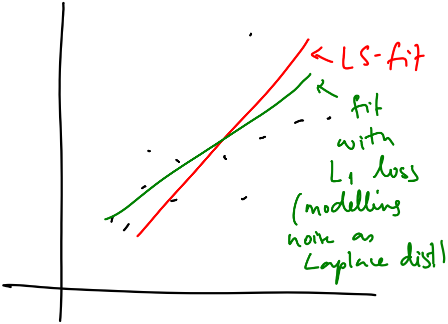
Can give confidence in our prediction as well.

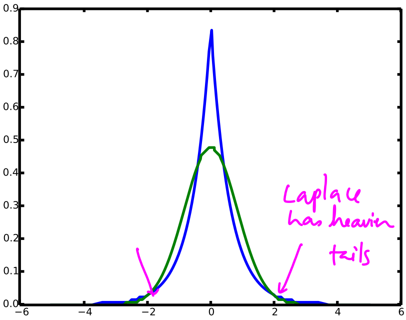
# Outliers and Laplace Distribution

With outliers least squares (and hence MLE with Gaussian model) can be quite bad.

Instead, we can model the noise (or uncertainty) in $y$ as a Laplace distribution

$$p(y \mid \mathbf{x}, \mathbf{w}, b) = \frac{1}{2b} \exp\left(-\frac{|y - \mathbf{x}^T\mathbf{w}|}{b}\right)$$



# Lookahead: Binary Classification

Bernoulli random variable $X$ takes value in $\{0, 1\}$. We parametrize using $\theta \in [0, 1]$.

$$p(1 \mid \theta) = \theta$$
$$p(0 \mid \theta) = 1 - \theta$$

More succinctly, we can write

$$p(x \mid \theta) = \theta^x (1 - \theta)^{1-x}$$

For classification, we will design models with parameter $\mathbf{w}$ that given input $\mathbf{x}$ produce a value in $f(\mathbf{x}; \mathbf{w}) \in [0, 1]$. Then, we can model the (binary) class labels as:

$$y \sim \mathrm{Bernoulli}(f(\mathbf{x}; \mathbf{w}))$$

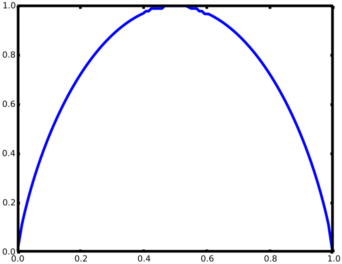
# Entropy

In information theory, entropy $H$ is a measure of uncertainty associated with a random variable.

$$H(X) = -\sum_x p(x)\log(p(x))$$

In the case of bernoulli variables (with parameter $\theta$) we get:

$$H(X) = -\theta\log(\theta) - (1-\theta)\log(1-\theta)$$



# Maximum Likelihood and KL-Divergence

Suppose we get data $x_1, \dots x_m$ from some unknown distribution $q$.

Attempt to find parameters $\theta$ for a family of distributions that best explains the data

$$
\begin{aligned}
\hat{\theta} &= \operatorname*{argmax}_{\theta} \prod_{i=1}^{m} p(x_i \mid \theta) \\
&= \operatorname*{argmax}_{\theta} \sum_{i=1}^{m} \log(p(x_i \mid \theta)) \\
&= \operatorname*{argmax}_{\theta} \frac{1}{m} \sum_{i=1}^{m} \log(p(x_i \mid \theta)) - \frac{1}{m} \sum_{i=1}^{m} \log(q(x_i)) \\
&= \operatorname*{argmin}_{\theta} \frac{1}{m} \sum_{i=1}^{m} \log\left(\frac{q(x_i)}{p(x_i \mid \theta)}\right) \\
&\to \operatorname*{argmin}_{\theta} \int \log\left(\frac{q(x)}{p(x)}\right) q(x)dx = \mathrm{KL}(q\|p)
\end{aligned}
$$

# Kullback-Leibler Divergence

KL-Divergence is “like” a distance between distributions

$$\mathrm{KL}(q\|p) = \sum_i \log \frac{q(x_i)}{p(x_i)} q(x_i) dx$$

$\mathrm{KL}(q\|q) = 0$

$\mathrm{KL}(q\|p) \geq 0$ for all distributions $p$