Causal Inference
CS 477-677

Learning From Data With Statistical Models

Ilya Shpitser

Outline

Uncertainty, And Random Variables

- Want to an answer questions in an uncertain world:
  - Will it rain tomorrow?
  - Who will win the US presidential election in 2020?
  - How likely am I to get lung cancer if I smoke?
- Will use random variables to deal with uncertainty.
- Example: "it rained today" (yes/no), call it X_1; "status of grass on my lawn" (wet/dry), call it X_2. Other examples?
- Will use probability (a number between 0 and 1) to encode uncertainty.
- Probability 1 means event is certain, 0 means event is impossible, somewhere between 0 and 1 quantifies how likely an event is.
- Write $p(X_1 = yes)$ to mean "probability it will rain today."

Joint, Conditional, and Marginal Distributions

- Will talk about probabilities of a single event, of multiple events, probability of one set of events given that we observed another:

$$p(\text{it rained}) = p(X_1 = \textit{yes}) = 1/4$$
$$p(\text{it rained and the lawn is wet}) = p(X_1 = \textit{yes}, X_2 = \textit{wet}) = 9/40$$
$$p(\text{the lawn is wet given that it rained}) = p(X_2 = \textit{wet} \mid X_1 = \textit{yes}) = 9/10$$

- A probability distribution maps event combinations to probabilities (think of a table of numbers that sum to 1).
- A conditional distribution maps event combinations to probability distributions (which are themselves maps).

X_1	X_2	$p(X_1, X_2)$	$p(X_2 \mid X_1) = p(X_1, X_2)/p(X_1)$	$p(X_1) = \sum_{x_2} p(X_1, X_2 = x_2)$
yes	*wet*	9/40	9/10	1/4
yes	*dry*	1/40	1/10	1/4
no	*wet*	3/20	1/5	3/4
no	*dry*	3/5	4/5	3/4

Densities

- For a real valued (continuous) random variable X, define the probability density function $f(x)$ to be

$$f(x) = \lim_{h \to 0} p(X \in (x, x + h)).$$

- $f(x)$ is the limit of the probability X has value in a little line segment around every point.
- For a set of real valued $\vec{X}$, can define the joint density $f(\vec{x})$ to be

$$f(\vec{x}) = \lim_{\vec{h} \to \vec{0}} p(\vec{X} \in (\vec{x}, \vec{x} + \vec{h})).$$

- $f(\vec{x})$ is the limit of the probability $\vec{X}$ has values in a little hypercube around every point $\vec{x}$.
- Examples: uniform for $x \in [0, 1]$, Gaussian.

Properties of Distributions, and Densities

- $p(\vec{X}) \geq 0$, $f(\vec{x}) \geq 0$. (non-negativity).
- $\sum_{\vec{x}} p(\vec{X} = \vec{x}) = 1$, $\int f(\vec{x}) d\vec{x} = 1$. (normalization).
- Note: $p(\vec{X}) \leq 1$, same is not true for densities!
  You can have $f(\vec{x}) > 1$!
- $p(X_1, \ldots X_k) = \prod_{i=1}^{k} p(X_i \mid X_{i-1}, \ldots X_1)$ (chain rule). Example:

$$\begin{aligned} p(X_3, X_2, X_1) &= p(X_3 \mid X_2, X_1) p(X_2 \mid X_1) p(X_1) \\ &= p(X_1 \mid X_2, X_3) p(X_2 \mid X_3) p(X_3). \end{aligned}$$

- $p(X_1 \mid X_2) = p(X_2 \mid X_1) p(X_1) / p(X_2)$ (Bayes rule).

Expected Value and Variance

- Expected value of numeric X ($E[X]$): average weighted by probabilities. Example:

$$X : \text{coin}, E[X] = 0 \cdot 0.5 + 1 \cdot 0.5 = 0.5.$$
$$X : \text{1d6 die}, E[X] = 1 \cdot 1/6 + \ldots + 6 \cdot 1/6 = 3.5.$$

- For densities:

$$E[X] = \int x \cdot f(x) dx$$

- Variance of X: $E[(X - E[X])^2]$. Example:

$$X : \text{coin}, Var[X] = (0 - 0.5)^2 \cdot 0.5 + (1 - 0.5)^2 \cdot 0.5 = 0.25$$
$$X : \text{1d6 die}, Var[X] = \sum_{i=1}^{6} (i - 3.5)^2/6 \approx 2.92$$

- $E[X]$: "average value of X," $Var[X]$: "spread of X."

Marginal and Conditional Independence

- Random variables A, B are marginally independent if $p(A, B) = p(A) \cdot p(B)$. Written $A \perp\!\!\!\perp B$.
- Example: two fair coins.
- Random variables A, B are conditionally independent given C if

$$p(A \mid B, C) = p(A \mid C)$$
$$p(B \mid A, C) = p(B \mid C)$$
$$p(A, B \mid C) = p(A \mid C) \cdot p(B \mid C)$$

- Written $A \perp\!\!\!\perp B \mid C$.
- Easily generalizes to *sets* of variables, too!
- Properties of conditional independence (semi-graphoid axioms):

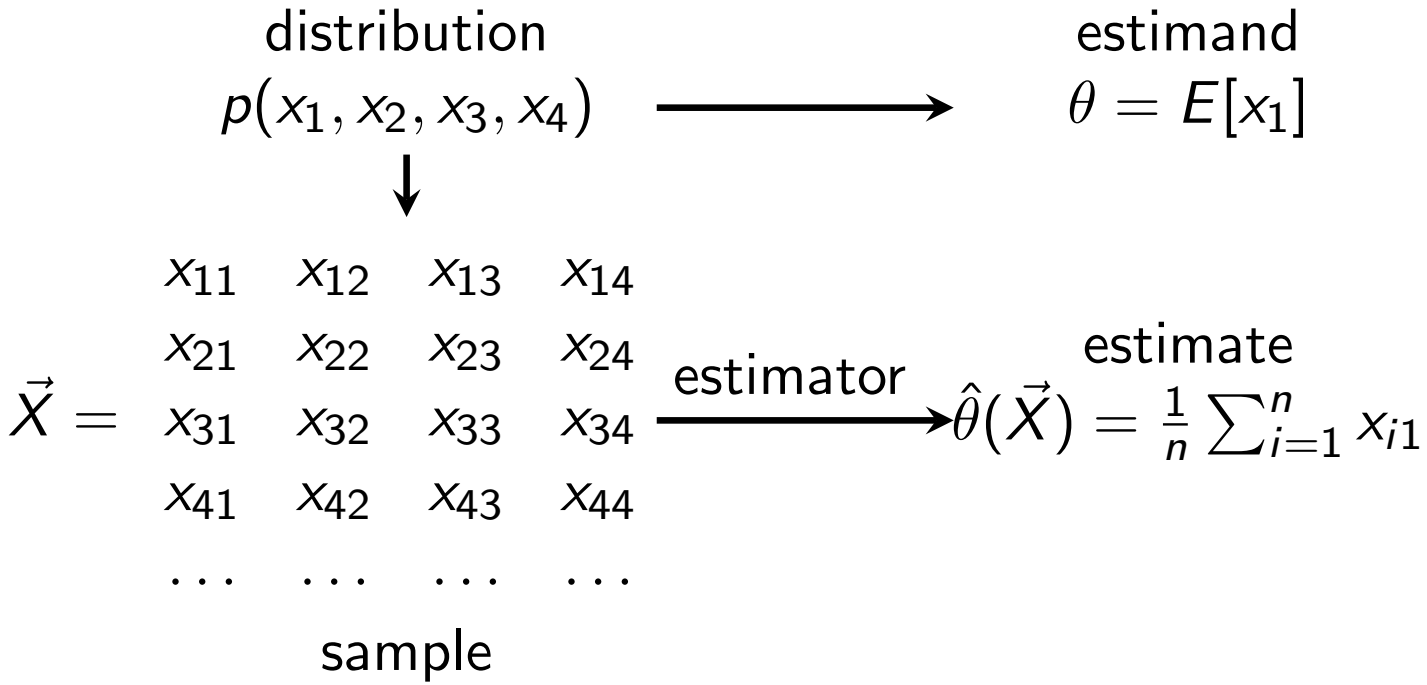
$$\vec{A} \perp\!\!\!\perp \vec{B} \mid \vec{C} \Leftrightarrow \vec{B} \perp\!\!\!\perp \vec{A} \mid \vec{C} \text{ (symmetry)}$$
$$\vec{A} \perp\!\!\!\perp \vec{B} \cup \vec{D} \mid \vec{C} \Leftrightarrow \vec{A} \perp\!\!\!\perp \vec{D} \mid \vec{C} \text{ and } \vec{A} \perp\!\!\!\perp \vec{B} \mid \vec{C} \cup \vec{D} \text{ (chain rule)}$$

- Warning: no finite axiomatization!

(Frequentist) Statistics View of Life

- A distribution $p(x_1, \ldots, x_k)$ captures properties of a population (college age people in the USA), or an object (a coin we flip).
- Some parameter or estimand of interest, say $E[X_1]$, or $p(X_2 = \text{heads})$.
- This is a fixed, but unknown property of the world.
- Can't measure everyone (too expensive), can't flip coin forever (too long), but can get a representative subset called a sample.
- Find an algorithm called an estimator to approximate parameter via an estimate.
- An estimate is a function of the sample, and so is a random variable, because the sample is random.

Graphical View



- Will represent the sample as a matrix, will often write as $\vec{X}$ or $\vec{X}_{n\times k}$.
- Columns: random variables in p, rows: draws from p.
- Will assume all rows are independent, identically distributed.
- Think coin flips.
- If we have time, may depart from this at the end of class.

What We Want From Estimators

- Accuracy: if $E[X_1] = 5$, want estimators to produce numbers that are roughly 5, not roughly -100.
- Dart analogy: want to hit around the bullseye on average.
- Precision: estimator output should not wildly fluctuate with different inputs.
- Dart analogy: want consistency, so our darts cluster around a small area, even if our aim is "off."
- Robustness: if we have weird or missing values in sample, or assumptions are wrong, don't output crazy things.
- Reliability: if we give estimator a range of different inputs, it still produces output reasonably close to correct most of the time.
- Trade offs involved here, can't have everything!

Accuracy (Lack Of Bias)

- Want to quantify bias: 0 means "no bias."
- Should compare θ and $\hat{\theta}(\vec{X})$.

Accuracy (Lack Of Bias)

- Want to quantify bias: 0 means "no bias."
- Should compare θ and $\hat{\theta}(\vec{X})$.
- But there is no single $\hat{\theta}(\vec{X})$! Estimator maps different samples $\vec{X}$ to different $\hat{\theta}(\vec{X})$.
- But we can average across possible samples, to get:

$$\theta - E[\hat{\theta}(\vec{X})] = \theta - \sum_{\vec{x}} (\hat{\theta}(\vec{X})) \cdot p(\vec{X}).$$

Accuracy (Lack Of Bias)

- Want to quantify bias: 0 means "no bias."
- Should compare θ and $\hat{\theta}(\vec{X})$.
- But there is no single $\hat{\theta}(\vec{X})$! Estimator maps different samples $\vec{X}$ to different $\hat{\theta}(\vec{X})$.
- But we can average across possible samples, to get:

$$\theta - E[\hat{\theta}(\vec{X})] = \theta - \sum_{\vec{x}} (\hat{\theta}(\vec{X})) \cdot p(\vec{X}).$$

- Estimator is unbiased if $\theta - E[\hat{\theta}(\vec{X})] = 0$.
- Note: for any specific $\vec{X}$, $\theta - \hat{\theta}(\vec{X})$ (called the error) may not be 0!
- Causal inference is about dealing with selection and confounding biases (much more on this later).

Precision (Lack Of Variance)

- Will quantify precision as variance of $\hat{\theta}(\vec{X})$ (remember $\hat{\theta}(\vec{X})$ is a random variable):

$$Var(\hat{\theta}(\vec{X})) = E[(\hat{\theta}(\vec{X}) - E[\hat{\theta}(\vec{X})])^2].$$

- As before $E[.]$ is with respect to $p(\vec{x})$.

Precision (Lack Of Variance)

- Will quantify precision as variance of $\hat{\theta}(\vec{X})$ (remember $\hat{\theta}(\vec{X})$ is a random variable):

$$Var(\hat{\theta}(\vec{X})) = E[(\hat{\theta}(\vec{X}) - E[\hat{\theta}(\vec{X})])^2].$$

- As before $E[.]$ is with respect to $p(\vec{x})$.
- Note: true θ appears nowhere here.
- We measure how tightly clustered our "darts" are, not how close to the bullseye we are.

Quantifying Precision And Accuracy Together

- A way to measure accuracy and precision is mean squared error:

$$\text{MSE} = E[(\theta - \hat{\theta}(\vec{X}))^2]$$

- Literally: the mean ($E[.]$) of the error squared.
- Important decomposition: MSE $=$ bias^2 $+$ variance:

$$E[(\theta - \hat{\theta}(\vec{X}))^2] = (\theta - E[\hat{\theta}(\vec{X})])^2 + E[(\hat{\theta}(\vec{X}) - E[\hat{\theta}(\vec{X})])^2]$$

- An inaccurate but precise estimator could have smaller MSE than an accurate but imprecise estimator.
- If the estimator is unbiased, MSE $= Var(\hat{\theta}(\vec{X}))$.

A Stronger Type Of Accuracy

- Unbiasedness is a fairly weak property.
- Say we want to estimate probability of the coin landing heads from an odd number of coin tosses.
- If the coin is fair, an estimator that gives 1 if #heads $>$ #tails, and 0 if #heads $<$ #tails is unbiased, since

A Stronger Type Of Accuracy

- Unbiasedness is a fairly weak property.
- Say we want to estimate probability of the coin landing heads from an odd number of coin tosses.
- If the coin is fair, an estimator that gives 1 if #heads $>$ #tails, and 0 if #heads $<$ #tails is unbiased, since

$$\begin{aligned} 0.5 - E[\hat{\theta}(\vec{X})] &= 0.5 - 1 \cdot p(\#\text{heads} > \#\text{tail}) - 0 \cdot p(\#\text{heads} < \#\text{tails}) \\ &= 0.5 - 1 \cdot 0.5 = 0 \end{aligned}$$

- For any given sequence of flips, the estimator is nowhere close!

A Stronger Type Of Accuracy

- Unbiasedness is a fairly weak property.
- Say we want to estimate probability of the coin landing heads from an odd number of coin tosses.
- If the coin is fair, an estimator that gives 1 if #heads $>$ #tails, and 0 if #heads $<$ #tails is unbiased, since

$$\begin{aligned} 0.5 - E[\hat{\theta}(\vec{X})] &= 0.5 - 1 \cdot p(\#\text{heads} > \#\text{tail}) - 0 \cdot p(\#\text{heads} < \#\text{tails}) \\ &= 0.5 - 1 \cdot 0.5 = 0 \end{aligned}$$

- For any given sequence of flips, the estimator is nowhere close!
- Want a property that states that as we get more and more data, we get closer and closer to true θ.

Consistency

- "More and more data" means a sequence of samples with more and more rows.
- Each sample has its own estimator, so we are really talking about a *sequence* of estimators.
- Such a sequence is consistent if as sample size grows to infinity, we are close to true θ with probability approaching 1.
- Formally, for any tiny ϵ,

$$\lim_{n \to \infty} p(|\hat{\theta}(\vec{X}_{n \times k}) - \theta| < \epsilon) = 1.$$

Consistency

- "More and more data" means a sequence of samples with more and more rows.
- Each sample has its own estimator, so we are really talking about a *sequence* of estimators.
- Such a sequence is consistent if as sample size grows to infinity, we are close to true θ with probability approaching 1.
- Formally, for any tiny ϵ,

$$\lim_{n \to \infty} p(|\hat{\theta}(\vec{X}_{n \times k}) - \theta| < \epsilon) = 1.$$

On the board

Consistency (cont.)

- Consistency is often a desirable property, but sometimes is sacrificed for other properties.
- Important: consistency is about long term behavior with lots of samples.
- A rule for constructing a consistent sequence of estimators is not guaranteed to behave well for a particular finite sample size.

Consistency (cont.)

- Consistency is often a desirable property, but sometimes is sacrificed for other properties.
- Important: consistency is about long term behavior with lots of samples.
- A rule for constructing a consistent sequence of estimators is not guaranteed to behave well for a particular finite sample size.
- Analogy for computer scientists: an algorithm may run in polynomial time (in P), but may have large "constant factors" that make it relatively inefficient at a particular input size.

Constructing Estimators (Example)

- Call a coin random variable X.
- Our data is 8 coin flips from X ($1 =$ heads, $0 =$ tails):

$$\vec{X}_{8\times 1} = (1, 0, 0, 0, 1, 1, 0, 0)^T.$$

- Want to model $p(X = 1)$ as q, and $p(X = 0)$ as $1 - q$. Then

$$p(X = x) = q^x \cdot (1 - q)^{1-x}$$

- This is called the *Bernoulli distribution* ($p(X) \sim \text{Bern}(q)$).

Constructing Estimators (Example)

- Call a coin random variable X.
- Our data is 8 coin flips from X ($1 =$ heads, $0 =$ tails):

$$\vec{X}_{8\times 1} = (1, 0, 0, 0, 1, 1, 0, 0)^T.$$

- Want to model $p(X = 1)$ as q, and $p(X = 0)$ as $1 - q$. Then

$$p(X = x) = q^x \cdot (1 - q)^{1-x}$$

- This is called the *Bernoulli distribution* ($p(X) \sim \text{Bern}(q)$).
- Intuitively, we should estimate $\hat{q} = 3/8$ from this data.
- How to think about this in general?

General Setup For Learning From Data

- Want to pick a "good" parameter value.
- We want a function that will tell us, for any parameter value, how surprised we should be to see the data.
- If surprised, parameter seems like a bad choice.
- If not surprised, parameter seems like a good choice.

General Setup For Learning From Data

- Want to pick a "good" parameter value.
- We want a function that will tell us, for any parameter value, how surprised we should be to see the data.
- If surprised, parameter seems like a bad choice.
- If not surprised, parameter seems like a good choice.
- Recall: rows in a sample are *independent*, so define this measure of surprise:

$$\mathcal{L}(\vec{X}_{8\times 1}; q) = \prod_{i=1}^{8} p(X = x_{i1}) = q^3 \cdot (1-q)^5.$$

- $\mathcal{L}(\vec{X}_{8\times 1}; q)$ is called the *likelihood function*.
- Want to pick q to maximize $\mathcal{L}(\vec{X}_{8\times 1}; q)$, or "minimize surprise."
- How to do this? Differentiate $\mathcal{L}$ with respect to q, set to 0.

The Likelihood Function And Learning Parameters

- Have a sample $\vec{X}_{n\times k}$, rows are data points, columns are features.
- Random variables $X_1, \ldots X_k$.
- Statistical model: $\{p(X_1, X_2, \ldots X_k; \pi) \mid \text{ some } \pi\}$ relating random variables.
- π is an unknown set of *parameters*.
- Assume data is independent, samples from some $p(X_1, X_2, \ldots X_k; \pi)$ in the set.
- Want to learn what we can about π.
- If $|\pi|$ is finite and independent of n, model is *parametric*.
- If $|\pi|$ is a function that scales with n, model is *non-parametric*.
- A part of p may be parametric, and a part non-parametric, in which case the model is *semi-parametric*.
- Statistical models are called probabilistic hypothesis classes in ML.

Maximizing Likelihood

- Define
$$\mathcal{L}(\vec{X}_{n\times k};\pi) = \prod_{i=1}^{n} p(X_1 = x_{i1}, \ldots X_k = x_{ik};\pi)$$
- Will often be easier to work with $\log \mathcal{L}$ (log-likelihood).
- Does not affect maximization but easier to work with.
- Solve for
$$\frac{\partial \log \mathcal{L}(\vec{X}_{n\times k};\pi)}{\partial \pi} = 0.$$
- Lots of issues here.

Issues with Maximizing Likelihood

- Equations of this type are known as estimating equations:

$$\frac{\partial \log \mathcal{L}(D_{n\times k}; \pi)}{\partial \pi} = 0.$$

- Problem 1: may not be solvable in closed form (transcendental equation).
- Iterative algorithms:
  - Grid search
  - Local search
- Problem 2: derivative at 0 might give local maximum. Or saddle point. Or minimum.
- Important class of functions where this does not happen.

Convex Functions

- Single global minimum if $(-\log \mathcal{L})$ is *convex* in π. A function $f(x)$ is convex if

$$f(t \cdot x_1 + (1-t) \cdot x_2) \leq t \cdot f(x_1) + (1-t) \cdot f(x_2).$$

- If function is convex, local search will often work very well (can we think of an example where it will not?)
- If the problem is non-convex, life gets hard...
- Function optimization is an interesting area, will leave aside for now.
- Another problem with maximizing likelihood, to be discussed later.

The Bernoulli (Biased Coin Flip) Model

- Data: a sequence of n binary outcomes $\vec{X}_{n\times 1}$ ("biased coin flips").
- Model:
$$p(X = x) = q^x \cdot (1 - q)^{1-x}.$$
- Likelihood:
$$\mathcal{L}(\vec{X}_{n\times 1}; \{q\}) = \prod_{i=1}^{n} q^{x_{i1}}(1 - q)^{1-x_{i1}}.$$
- Log likelihood:
$$\log \mathcal{L}(\vec{X}_{n\times 1}; \{q\}) = \sum_{i=1}^{n} x_{i1} \log q + (1 - x_{i1}) \log(1 - q).$$
- Maximum likelihood estimator for q: $\hat{q}(\vec{X}_{n\times 1}) = \frac{1}{n} \sum_{i=1}^{n} x_{i1}$.

The Bernoulli (Biased Coin Flip) Model

- Data: a sequence of n binary outcomes $\vec{X}_{n\times 1}$ ("biased coin flips").
- Model:
$$p(X = x) = q^x \cdot (1 - q)^{1-x}.$$
- Likelihood:
$$\mathcal{L}(\vec{X}_{n\times 1}; \{q\}) = \prod_{i=1}^{n} q^{x_{i1}}(1 - q)^{1-x_{i1}}.$$
- Log likelihood:
$$\log \mathcal{L}(\vec{X}_{n\times 1}; \{q\}) = \sum_{i=1}^{n} x_{i1} \log q + (1 - x_{i1}) \log(1 - q).$$
- Maximum likelihood estimator for q: $\hat{q}(\vec{X}_{n\times 1}) = \frac{1}{n}\sum_{i=1}^{n} x_{i1}$.

On the board

The Gaussian (Normal) Density

One of the most important densities in statistics:

$$f(x; \{\mu, \sigma^2\}) = \mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

- If $X \sim \mathcal{N}(\mu, \sigma^2)$, $E[X] = \mu$, $Var[X] = \sigma^2$.
- Lots of continuous valued variables in Nature are approximately Gaussian.
- Why? Central limit theorem.
- Averages (and sums) of lots of random variables (under some conditions) will be Gaussian.
- Height, test performance, etc.
- Important in statistical theory, will return to this later.

The Gaussian Model

- Data: a sequence of n realizations of X, written as $\vec{X}_{n\times 1}$.
- Model:

$$f(X = x) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

- Likelihood (density as "measure of surprise"):

$$\mathcal{L}(\vec{X}_{n\times 1}; \{\mu, \sigma^2\}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left\{-\frac{(x_{i1}-\mu)^2}{2\sigma^2}\right\}$$

- Log likelihood:

$$\log \mathcal{L}(\vec{X}_{n\times 1}; \{\mu, \sigma^2\}) = \sum_{i=1}^{n} \left\{-\frac{(x_{i1}-\mu)^2}{2\sigma^2}\right\} - n \log \sqrt{2\sigma^2\pi}.$$

The Gaussian Model

- Data: a sequence of n realizations of X, written as $\vec{X}_{n\times 1}$.
- Model:

$$f(X = x) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

- Likelihood (density as "measure of surprise"):

$$\mathcal{L}(\vec{X}_{n\times 1}; \{\mu, \sigma^2\}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left\{-\frac{(x_{i1}-\mu)^2}{2\sigma^2}\right\}$$

- Log likelihood:

$$\log \mathcal{L}(\vec{X}_{n\times 1}; \{\mu, \sigma^2\}) = \sum_{i=1}^{n} \left\{-\frac{(x_{i1}-\mu)^2}{2\sigma^2}\right\} - n \log \sqrt{2\sigma^2\pi}.$$

- Maximum likelihood estimators for μ, σ^2:

On the board

The Gaussian Model

- Data: a sequence of n realizations of X, written as $\vec{X}_{n\times 1}$.
- Model:

$$f(X = x) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

- Likelihood (density as "measure of surprise"):

$$\mathcal{L}(\vec{X}_{n\times 1}; \{\mu, \sigma^2\}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left\{-\frac{(x_{i1}-\mu)^2}{2\sigma^2}\right\}$$

- Log likelihood:

$$\log \mathcal{L}(\vec{X}_{n\times 1}; \{\mu, \sigma^2\}) = \sum_{i=1}^{n} \left\{-\frac{(x_{i1}-\mu)^2}{2\sigma^2}\right\} - n \log \sqrt{2\sigma^2\pi}.$$

- Maximum likelihood estimators for μ, σ^2:

$$\hat{\mu}(\vec{X}_{n\times 1}) = \frac{1}{n}\sum_{i=1}^{n} x_{i1}; \qquad \widehat{\sigma^2}(\vec{X}_{n\times 1}) = \frac{1}{n}\sum_{i=1}^{n}(x_{i1} - \hat{\mu}(\vec{X}_{n\times 1}))^2$$

Next time: Linear and Logistic Regression Models