ISTA 421/521
Introduction to Machine Learning

Lecture 5: Probability Review, Expectation, Distributions

Clayton T. Morrison
clayton@sista.arizona.edu
Gould-Simpson 819
Phone 621-6609

Special Thanks to Rev. Dawson

9 September 2014

Next Topics

- Probability Basics
- Expectation
- Continuous probability
- Distributions
- Likelihood

Least Squares (Linear) Regression

- Model t as a *linear* function of $x_1, x_2, \ldots$.
- Choose the "best" model, of the form

$$\hat{t} = \hat{f}(\mathbf{x}; \mathbf{w}) = w_0 + w_1 x_1 + \cdots + w_K x_K \tag{1}$$

- "Best": select $\mathbf{w}$ to minimize the *loss*

$$\mathcal{L}(\mathbf{w}; \mathbf{X}, \mathbf{t}) = \sum_{n=1}^{N} (t_n - \hat{f}(\mathbf{x}_n; \mathbf{w}))^2 \tag{2}$$

- Can generalize to non-linear models, but principle is the same: pick the function $\hat{f}$ in a particular *function class*, $\mathcal{F}$ that "minimizes badness".

Why Squared Error?

- Squared error loss has a natural geometric definition that gives a vector of "retrodictions", $\hat{\mathbf{t}} = (\hat{t}_1, \ldots, \hat{t}_N)$ that is as close as possible to the vector of observations $\mathbf{t} = (t_1, \ldots, t_N)$ while respecting the linear constraint.
- "Closeness" measured by the usual Euclidean distance between two points in a (K-dimensional) vector space

$$|\mathbf{u} - \mathbf{v}|^2 = \sum_{k=1}^{K} (u_k - v_k)^2 \tag{3}$$

(in 2D this is the Pythagorean theorem)

Okay, But Why *Really*?

Short Answers:

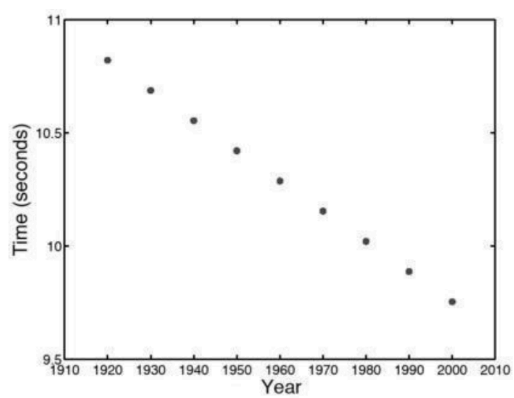
- It "feels right".
- It often works well in practice.

Deeper answer:

- it is the **maximum likelihood** solution under a natural **probabilistic generative model**.

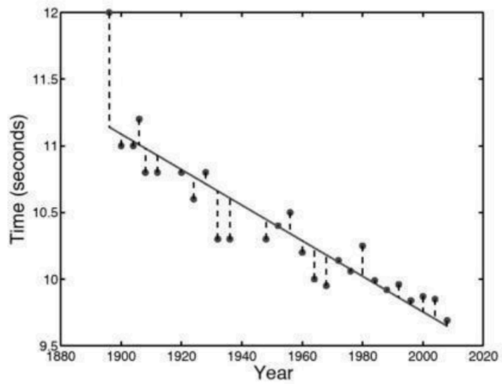
Maximum What? Generative What?

Our original formulation of the model was **deterministic**: for a given $\mathbf{x}$, the model yields the same t every time.

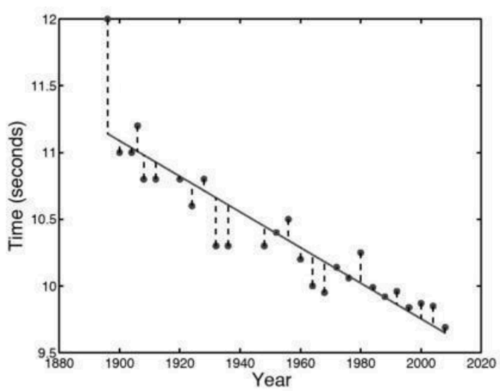


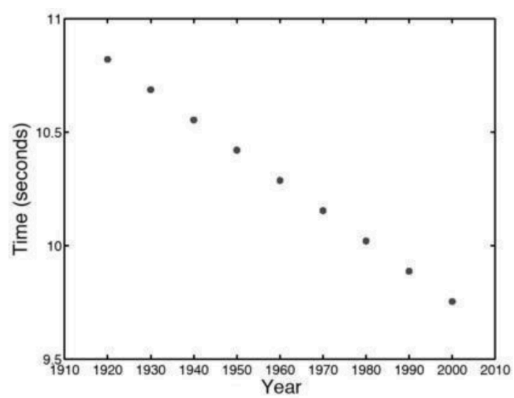
A More Realistic Model

In most interesting applications, life is more complicated.



Think Generatively





Data generated from linear model

Want degree of confidence in predictions and parameter values

Adding Error to the Model

- We capture this added complexity with a "catchall" error term, ε.

$$t = w_0 + w_1 x_1 + \cdots + w_k x_k + \varepsilon \tag{4}$$

- ε is sometimes positive, sometimes negative, and can be different for two cases even if all their x values are the same.
- It is a different beast from the variables x, w and t: it is a **random variable**.
- Captures all the factors that we are not modeling.

Sample Space

Sample Space

A **sample space**, S, is

1. Classical/objectivist defintion: a collection of possible **outcomes** of a **random experiment** (The coin will come up heads or tails. The die will come up 1,2,3,4,5 or 6.)
2. Bayesian/subjectivist definition: a collection of "possible worlds" that we might be in (The coin has come up heads or tails. The cat is alive or dead.)

When needed, we denote a generic individual outcome by ω, and can say, e.g., "for each $\omega \in S$, ..."

Events

Event

An **event** is a *subset* of the sample space that does or does not contain (is true or false for) a particular outcome/possible world.

- The coin comes up heads.
- The cat is alive.
- The die shows an even number.

Semantics of Set Operations

Equivalence between "set" and "proposition" representations.

1. Set E: outcomes s.t. proposition E is true.
2. Union, $E \cup F$: logical OR between propositions E and F.
3. Intersection, $E \cap F$: logical AND
4. Complement, E^C: logical negation

Probability Space

Probability Space

A **probability space** is a sample space, S, augmented with a function, P, that assigns a **probability** to each event, $E \subset S$.

Kolmogorov Axioms

1. $0 \leq P(E) \leq 1$ for all $E \subset S$.
2. $P(S) = 1$.
3. If $E \cap F = \varnothing$ then $P(E \cup F) = P(E) + P(F)$.

Important Consequences

1. $P(\varnothing) = 0$.
2. $P(E^C) = 1 - P(E)$
3. In general, $P(E \cup F) = P(E) + P(F) - P(E \cap F)$.

Random Variable

Random Variable

- Formally, a **random variable** is a function, X, that assigns a number to each outcome in S (e.g., dead $\to 0$, alive $\to 1$).
- Key consequence: a random variable divides the sample space into **equivalence classes**: sets of outcomes that share some property (differ only in ways irrelevant to X)

Example

- Let $S =$ all sequences of 3 coin tosses.
- We can define a r.v. X that counts number of heads.
- Then HHT and HTH are equivalent in the eyes of X:

$$X(HHT) = X(HTH) = 2$$