ECON 557 – Advanced Data Analysis

Michael T. Sandfort

Department of Economics
Masters in Applied Economics Program
Georgetown University

January 20, 2023

What are statistical models "for"?

- Description
- Prediction
- Causal Analysis

- The focus in this course will be on prediction.

Notation

Start by fixing some notation:

- $\mathcal{X}$ is the covariate space, or feature space, ("input space" is also popular) with typical element $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_m \end{bmatrix}$.
- $\mathcal{Y}$ is the output space, or response space ("label space" is also popular) with typical element y.

- Example: $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, (familiar from single-variable linear regression).
- Example: $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{0, 1\}$, (familiar from single-variable logit regression).

Notation

- $F_{\mathbf{x},y}$ is a joint probability distribution on $\mathcal{X} \times \mathcal{Y}$.
- $F_{\mathbf{x}}$ is the marginal distribution on $\mathbf{x}$ (i.e., $F_{\mathbf{x},y}$ integrated over y).
- $F_{y|\mathbf{x}}$ is the conditional distribution of y given $\mathbf{x} \in \mathcal{X}$.
- Expectations, variances and other moments relative to F (or its marginals or conditionals) are taken in the usual way:

$$\mathbb{E}[h(\mathbf{x}, y)] \equiv \int h dF_{\mathbf{x},y} = \int h f_{\mathbf{x},y} d\mathbf{x} dy$$

where $f_{\mathbf{x},y}$ is the joint density (when it exists).

Refresher on Probability Distributions F

- Univariate populations
  - Bernoulli
  - Normal
  - Chi-Square
  - Cauchy
- Bivariate populations
  - Bivariate Bernoulli
  - Bivariate Normal

Univariate Populations: Bernoulli

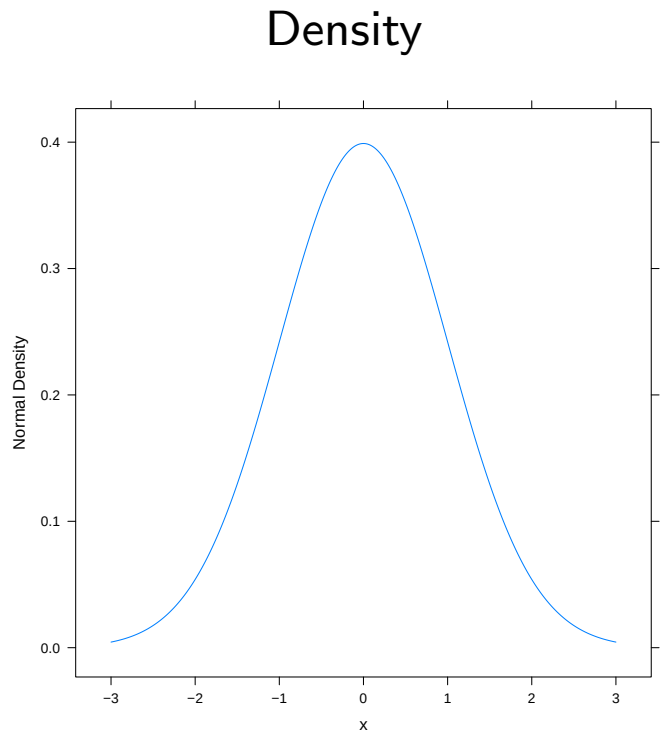
If $y \sim$ Bernoulli(p), then

- The sample space for y is $\mathcal{Y} = \{0, 1\}$;
- $y = 1$ with probability p, $y = 0$ with probability $1 - p$;
- Mean: $\mathbb{E}[y] \equiv \mu = \mathbb{P}(y = 0) \cdot 0 + \mathbb{P}(y = 1) \cdot 1 = p$
- Variance:
  $\mathbb{V}[y] \equiv \sigma^2 = \mathbb{E}[y^2] - \mathbb{E}^2[y] = ((1 - p) \cdot 0^2 + p \cdot 1^2) - p^2 = p(1 - p)$
- Standard Error: $\sqrt{\mathbb{V}[y]} = \sigma = \sqrt{p(1 - p)}$

Univariate Populations: Normal

If $y \sim \text{Normal}(0, 1)$, then

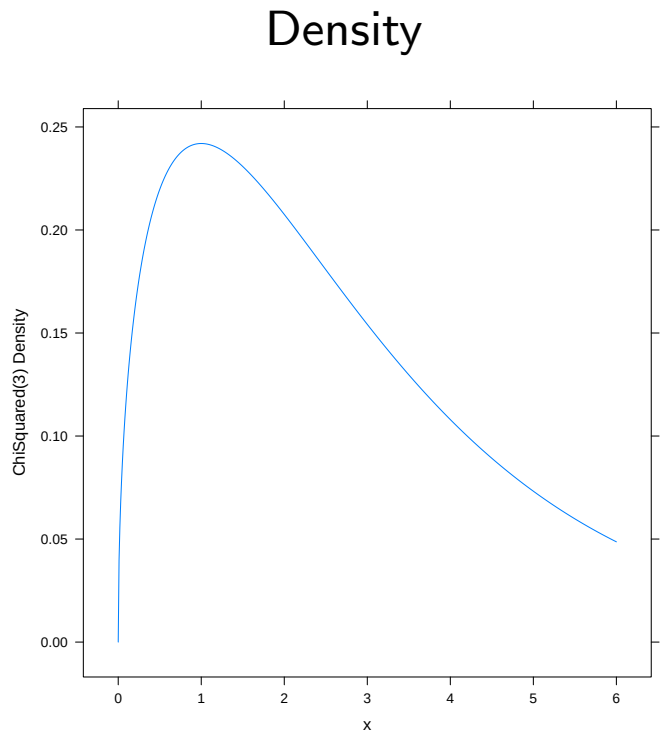
- Sample Space: $\mathcal{Y} = \mathbb{R}$
- Mean: $\mathbb{E}[y] \equiv \mu = 0$
- Variance: $\mathbb{V}[y] \equiv \sigma^2 = 1$
- Standard Error: $\sqrt{\mathbb{V}[y]} \equiv \sigma = 1$



Univariate Populations: χ^2

If $y \sim \chi^2(3)$ then,

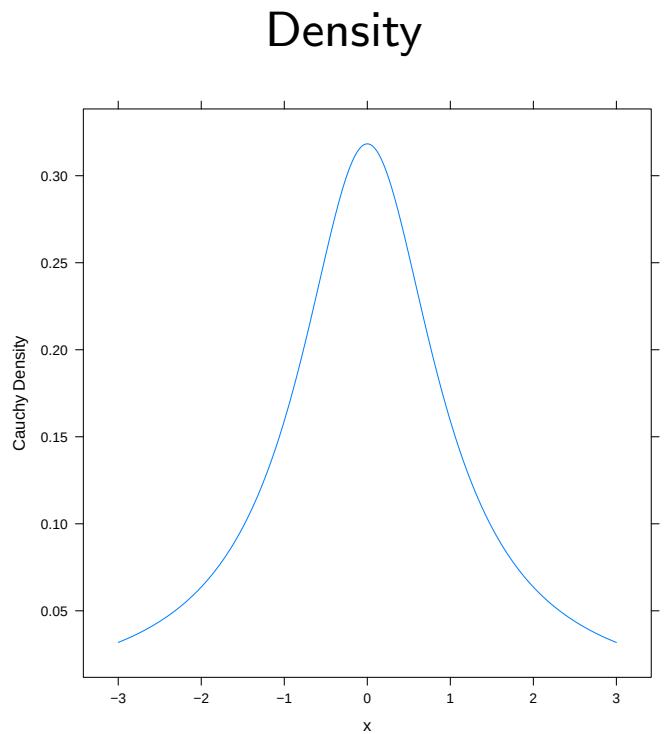
- Sample Space: $\mathcal{Y} = \mathbb{R}^+$
- Mean: $\mathbb{E}[y] \equiv \mu = 3$
- Variance: $\mathbb{V}[y] \equiv \sigma^2 = 6$
- Standard Error: $\sqrt{\mathbb{V}[y]} \equiv \sigma = \sqrt{6}$



Univariate Populations: Cauchy

If $y \sim$ Cauchy then,

- Sample Space: $\mathcal{Y} = \mathbb{R}$
- Mean: $\mathbb{E}[y] \equiv \mu = ?$
- Variance: $\mathbb{V}[y] \equiv \sigma^2 = ?$
- Standard Error: $\sqrt{\mathbb{V}[y]} \equiv \sigma = ?$



Sidebar: Populations (DGPs) in R

- For any univariate population, there are typically four key items of interest:
  - The probability mass function $F : \mathcal{Y} \to [0, 1]$.
  - The quantile function $F^{-1} : [0, 1] \to \mathcal{Y}$.
  - The ability to draw a random sample $\{y_i\}_{i=1}^N$ from F.
  - Where it exists, the density function $\frac{dF}{dy} = f : \mathcal{Y} \to \mathbb{R}^+$.
- In R, all four of these functions are available for a wide variety of populations, following the naming convention:
  - `p<<name>>()` provides the [p]robability mass function.
  - `q<<name>>()` provides the [q]uantile function.
  - `r<<name>>()` generates a [r]andom sample.
  - `d<<name>>()` provides the [d]ensity function.

Sidebar: Populations (DGPs) in R

- A standard normal random variable will be less than $y = 1.96$ about 97.5% percent of the time:

```
> pnorm(q=1.96) # or pnorm(mean=0,sd=1,q=1.96)
[1] 0.9750021
```

- The mean of a $\chi^2(3)$ is 3. What is the median? Since the median is $F^{-1}(0.5)$:

```
> qchisq(p=.5,df=3)
[1] 2.365974
```

- Draw a sample of size $n = 10$ from a Bernoulli distribution with parameter $p = .25$:

```
> rbinom(n=10,size=1,prob=.25)
 [1] 0 0 0 0 0 1 0 0 0 0
```

- What is (approximately) the mode of a N(3,2)?

```
> xvec <- seq(0,6,by=.01)
> max_idx <- which.max(dnorm(x=xvec,mean=3,sd=sqrt(2)))
> xvec[max_idx]
[1] 3
```

Bivariate Populations: Joint Bernoulli

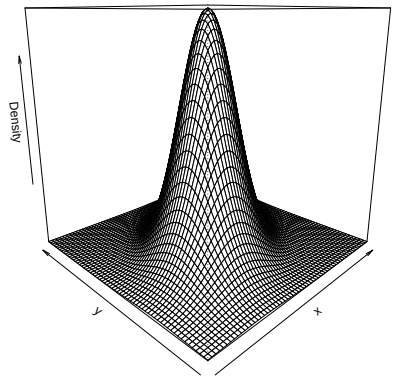
This is an exam favorite!

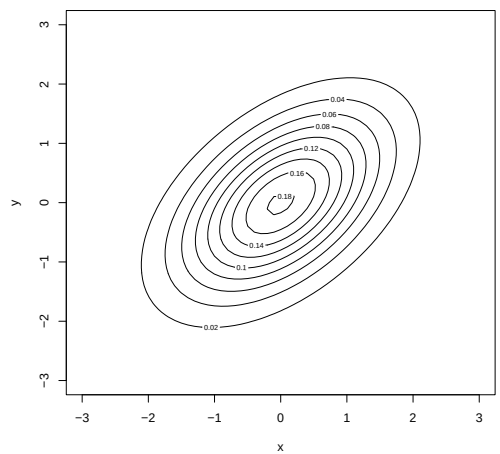
- Sample Space: $\mathcal{X} \times \mathcal{Y} = \{0, 1\} \times \{0, 1\}$
- Means: $\mu_x =?$ and $\mu_y =?$
- Variances: $\sigma_x^2 =?$ and $\sigma_y^2 =?$
- Standard Errors: $\sigma_x =?$ and $\sigma_y =?$
- Correlation Coefficient: $\rho =?$
- Marginal Distribution on $\mathbf{x}$: $F_{\mathbf{x}} =?$
- Conditional Distribution: $F_{y|\mathbf{x}} =?$

	$x = 0$	$x = 1$
$y = 0$	0.15	0.25
$y = 1$	0.35	0.25

Bivariate Populations: BVN(0,0,1,1,.5)

- Sample Space: $\mathcal{X} \times \mathcal{Y} = \mathbb{R}^2$
- Means: $\mu_x = 0$ and $\mu_y = 0$
- Variances: $\sigma_x^2 = 1$ and $\sigma_y^2 = 1$
- Standard Errors: $\sigma_x = 1$ and $\sigma_y = 1$
- Correlation Coefficient: $\rho = 0.5$
- Marginal Distribution on $\mathbf{x}$: $F_{\mathbf{x}} = ?$
- Conditional Distribution: $F_{y|\mathbf{x}} = ?$





Exponential Dispersion Families

Almost all of the distributions we have covered so far are in the exponential dispersion family of distributions.

- The binomial (which includes the Bernoulli), normal, and Poisson distributions are all in the exponential dispersion family.
- Any joint distribution on discrete random variables (e.g., the bivariate Bernoulli we'll see momentarily) is in the exponential dispersion family.[1]

In particular, any distribution for which we can describe the measure of probability mass as

$$f(y|\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right]$$

is in the exponential dispersion family.

- $\theta \in \mathbb{R}$ is the canonical parameter,
- $\phi > 0$ is the dispersion parameter,
- $\frac{b(\theta)}{\phi}$ is known as the log normalizer or cumulant.

[1] This isn't hard to show, but we won't do it here.

Exponential Families

- We will focus only on single-parameter exponential dispersion families.
- The normal distribution is actually a two-parameter exponential family, but the variance is commonly treated as a "nuisance parameter." Taking the variance as the dispersion makes the one-parameter normal distribution an exponential dispersion family for $\frac{y}{\sigma}$.
- The multinomial distribution is another common example of a multi-parameter exponential family distribution.
- When we talk about the number of parameters in an exponential dispersion family, that is the dimension of the canonical parameter space, not than the number of parameters in $\boldsymbol{\beta}$, as we will see next time.

The Normal is an Exponential Dispersion Family Distribution

The normal distribution with fixed σ^2 allocates mass over $y \in \mathbb{R}$ according to

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

Simplify this by expanding $(y-\mu)^2$, so

$$f(y) = \exp\left(\frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right)$$

Therefore the normal with fixed σ^2 is an exponential dispersion family with

- $\theta = \mu$
- $\phi = \sigma^2$
- $b(\theta) = \frac{1}{2}\mu^2 = \frac{1}{2}\theta^2$
- $c(y, \phi) = -\frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) = -\frac{y^2}{2\phi} - \frac{1}{2}\log(2\pi\phi)$

The Bernoulli is an Exponential Dispersion Family Distribution

The Bernoulli distribution allocates mass over $y \in \{0, 1\}$ according to

$$\begin{aligned} f(y) &= p^y (1-p)^{1-y} \\ &= \exp\left[y \log \frac{p}{1-p} + \log(1-p) \right] \end{aligned}$$

Therefore the Bernoulli is an exponential dispersion family with

- $\theta = \log \frac{p}{1-p}$ (the "log odds ratio")
- $\phi = 1$
- $b(\theta) = -\log(1-p) = \log(1+e^\theta)$
- $c(\theta, \phi) = 0$

The Poisson is an Exponential Dispersion Family Distribution

The Poisson distribution allocates mass over $y \in \mathbb{N}$ according to

$$\begin{aligned} f(y) &= \frac{\lambda^y e^{-\lambda}}{y!} \\ &= \exp\left[y \log \lambda - \lambda - \log(y!)\right] \end{aligned}$$

Therefore the Poisson is an exponential dispersion family with

- $\theta = \log \lambda$
- $\phi = 1$
- $b(\theta) = \lambda = e^{\theta}$
- $c(\theta, \phi) = -\log(y!)$

The Mean of Exponential Dispersion Family Distributions

A probability distribution must aggregate to one, so we have

$$\int f(y|\theta,\phi)dy = 1$$

when the measure on y is continuous, or a related condition when the measure on y is discrete.

Then using the definition of an exponential dispersion family

$$\begin{aligned}
&\int \exp\left[\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right] dy = 1 \\
\implies &\exp\left(-\frac{b(\theta)}{\phi}\right)\int \exp\left(\frac{y\theta}{\phi} + c(y,\phi)\right) dy = 1 \\
\implies &\frac{b(\theta)}{\phi} = \log\int \exp\left(\frac{y\theta}{\phi} + c(y,\phi)\right) dy
\end{aligned}$$

The Mean of Exponential Dispersion Family Distributions

Differentiating the expression from the previous page gives

$$\frac{b'(\theta)}{\phi} = \frac{\int \frac{y}{\phi} \exp\left(\frac{y\theta}{\phi} + c(y,\phi)\right) dy}{\int \exp\left(\frac{y\theta}{\phi} + c(y,\phi)\right) dy}$$

from which it follows that

$$\begin{aligned}
\frac{b'(\theta)}{\phi} &= \int \frac{y}{\phi} \frac{\exp\left(\frac{y\theta}{\phi} + c(y,\phi)\right)}{\exp\left(\frac{b(\theta)}{\phi}\right)} dy \\
&= \int \frac{y}{\phi} \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right) dy \\
&= \int \frac{y}{\phi} f(y|\theta,\phi) dy \\
&= \frac{1}{\phi} \mathbb{E}[y|\theta,\phi]
\end{aligned}$$

so $b'(\theta) = \mathbb{E}[y|\theta,\phi] \equiv \mu$, from which it follows that $\theta = (b')^{-1}(\mu)$.

The Variance of Exponential Dispersion Family Distributions

Starting again from the first step on the previous slide,

$$b'(\theta) = \exp\left(-\frac{b(\theta)}{\phi}\right)\int y \exp\left(\frac{y\theta}{\phi} + c(y,\phi)\right) dy$$

and differentiating again gives

$$\begin{aligned} b''(\theta) &= -\frac{b'(\theta)}{\phi}b'(\theta) + \exp\left(-\frac{b(\theta)}{\phi}\right)\int \frac{y^2}{\phi}\exp\left(\frac{y\theta}{\phi} + c(y,\phi)\right) dy \\ &= -\frac{\mu^2}{\phi} + \frac{1}{\phi}\mathbb{E}[y^2|\theta,\phi] \\ &= \frac{1}{\phi}\left\{-(\mathbb{E}[y|\theta,\phi])^2 + \mathbb{E}[y^2|\theta,\phi]\right\} = \frac{1}{\phi}\mathbb{V}[y|\theta,\phi] \end{aligned}$$

As before, we want to connect this to the expectation

$$\begin{aligned} \mathbb{V}[y|\theta,\phi] &= \phi b''(\theta) \\ &= \phi b''((b')^{-1}(\mu)) \\ &= \phi\mathcal{V}(\mu) \end{aligned}$$

where $\mathcal{V}(\mu)$ is called the variance function. Note that ϕ scales the variance, but does not affect the mean.

Mean and Variance of the Normal Distribution

For the normal distribution, we showed

- $\theta = \mu$
- $\phi = \sigma^2$
- $b(\theta) = \frac{1}{2}\theta^2$

Therefore,

$$\begin{aligned} \mu = b'(\theta) &= \theta \\ \mathcal{V}(\mu) = b''(\theta) &= 1 \end{aligned}$$

which shows that

$$\begin{aligned} \mathbb{E}[y|\theta, \phi] &= \theta &&= \mu \\ \mathbb{V}[y|\theta, \phi] &= \phi \cdot 1 &&= \phi \end{aligned}$$

Mean and Variance of the Bernoulli Distribution

For the Bernoulli distribution, we showed

- $\theta = \log\left(\frac{p}{1-p}\right)$
- $\phi = 1$
- $b(\theta) = \log(1 + e^{\theta})$

Therefore,

$$\mu = b'(\theta) = \frac{e^{\theta}}{1 + e^{\theta}}$$
$$\mathcal{V}(\mu) = b''(\theta) = \frac{e^{\theta}}{(1 + e^{\theta})^2}$$

which shows that

$$\mathbb{E}[y|\theta, \phi] = \frac{p}{1-p}\frac{1-p}{1} = p$$
$$\mathbb{V}[y|\theta, \phi] = \phi\frac{p}{1-p}\frac{(1-p)^2}{1} = p(1-p)$$

Mean and Variance of the Poisson Distribution

For the Poisson distribution, we showed

- $\theta = \log \lambda$
- $\phi = 1$
- $b(\theta) = e^{\theta}$

Therefore,

$$\begin{aligned} \mu = b'(\theta) &= e^{\theta} \\ \mathcal{V}(\mu) = b''(\theta) &= e^{\theta} \end{aligned}$$

which shows that

$$\begin{aligned} \mathbb{E}[y|\theta, \phi] &= \exp(\log \lambda) &&= \lambda \\ \mathbb{V}[y|\theta, \phi] &= \phi \exp(\log \lambda) &&= \lambda \end{aligned}$$

Prediction

- A predictor $h : \mathcal{X} \to \mathcal{Y}$ is just a mapping from the set of covariates $\mathcal{X}$ to the set of responses $\mathcal{Y}$.
- The idea is that if we see $\mathbf{x} \in \mathcal{X}$, but don't see the associated $y \in \mathcal{Y}$, we can use $\widehat{y} \equiv h(\mathbf{x})$, as a hypothetical value for y, also called a predicted response, or just prediction.
- Because it generates hypothetical values, the function $h(\cdot)$ is also often called a hypothesis.
- It bears repeating that $h(\cdot)$ is a function – it provides a hypothetical value $\widehat{y}$ which depends on what value of the covariate $\mathbf{x}$ we put in.
- To emphasize that $h(\cdot)$ is a function, it is sometimes called a prediction rule.

Evaluating Predictions and Predictors

- To test the quality of the **prediction** $h(\mathbf{x})$ for any population pair $(\mathbf{x}, y)$, it is natural to compare the true response y from the pair with the predicted response $\widehat{y} = h(\mathbf{x})$.
- In some settings, $y - \widehat{y}$ is called the prediction error; however, there are circumstances where such a treatment makes little or no sense (e.g., multinomial outcomes $\mathcal{Y} = \{1, 2, \ldots, J\}$).
- To test the quality of the **predictor** h, we want to compare $(\mathbf{x}, y)$ with $(\mathbf{x}, h(\mathbf{x}))$ throughout $\mathcal{X}$.

Best (?) Predictors

Even if $\mathcal{Y} = \mathbb{R}$, so computing $y - \widehat{y}$ might make sense, everyone might not agree about what makes a "good" prediction versus a "bad" one.

Does "best" mean...

- Minimize total (presumably, absolute) prediction error?
- Extra penalty for big errors? How much?
- Are positive errors worse than negative errors? Better?
- Getting it exactly right? That is, are all errors equally bad, regardless of direction or magnitude?

Best (?) Predictors

Perhaps what makes a "good" prediction even depends on the prediction environment.

Consider the case of a fingerprint scanner, where $\mathcal{X} = \mathbb{R}^d$ is a high-dimensional vector representing a fingerprint scan. $\mathcal{Y} = \{0, 1\}$ classifies whether the fingerprint belongs to you (1) or to an impostor (0).

- If the fingerprint scanner is attached to a point of sale (POS) device at a supermarket, then "Type I" errors ($\widehat{y} = 0$ when $y = 1$, i.e., rejecting your fingerprint) and "Type II" errors ($\widehat{y} = 1$ when $y = 0$, i.e., accepting the intruder's fingerprint) are both modestly costly (inconvenience vs. "shrinkage").
- If the fingerprint scanner is attached to a door at a SCIF (holding top secret documents), then the errors are highly asymmetric in their consequence: "Type I" errors are an employee inconvenience, while "Type II" errors could be devastating.

The Loss Function

- The loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ describes the consequences (in terms of cost or negative utility) of differences between true and predicted values.
- Common loss functions include:
  - Mean squared error (MSE) loss. For $\mathcal{Y} = \mathbb{R}$, let $\ell(\widehat{y}, y) \equiv (\widehat{y} - y)^2$.
  - Least absolute deviation (LAD) loss. For $\mathcal{Y} = \mathbb{R}$, let $\ell(\widehat{y}, y) \equiv |\widehat{y} - y|$.
  - Zero-one loss. For $\mathcal{Y} = \{0, 1\}$ or other $\mathcal{Y} \subset \mathbb{Z}$, $\ell(\widehat{y}, y) \equiv \mathbb{1}[\widehat{y} \neq y]$.

Risk

- The risk of a hypothesis $h(\cdot)$ is just the expected loss of using h, considered against the background of all possible true outcomes $(\mathbf{x}, y)$, or

$$R(h) \equiv \mathbb{E}_{\mathbf{x},y}[\ell(h(\mathbf{x}), y)] = \int \ell(h(\mathbf{x}), y) dF_{\mathbf{x},y}(\mathbf{x}, y)$$

  according to our earlier definition of expectation. Of course, computing the risk means we need to know $F_{\mathbf{x},y}$.

- Even though $\widehat{y}(\mathbf{x})$ has a "hat," don't confuse it with an estimate or estimator.

- We haven't even mentioned data yet – evaluating predictors is a population exercise.

Minimum-Risk Prediction

- Taking the risk function $R(h)$ from the previous slide as our objective, we are now in a position to describe best (minimum-risk) prediction.
- Let $\mathcal{H}$ be the set of all hypotheses we wish to consider.
- The Bayes' risk R^* is the minimum of the risk functional $R(\cdot)$ over all hypotheses $h \in \mathcal{H}$:

$$R^* \equiv \min_{h \in \mathcal{H}} R(h)$$

  and we'll call the minimizer h^* the Bayes optimal hypothesis.[2]
- Evidently, different loss functions $\ell(\cdot, \cdot)$ can give rise to different Bayes optimal hypotheses h^*.

[2] If you're mathematically sophisticated enough to know that the minimum above should actually be an infimum, you also know that a minimizer may not exist. Good for you!

Restricting our Search for the Best Predictor

- We can limit the universe $\mathcal{H}$ of hypotheses over which we optimize, as we will next show through a series of examples on the next few slides.
- Evidently, if $\mathcal{H}$ is a space which includes all functions $h : \mathcal{X} \to \mathcal{Y}$, while $\mathcal{H}_0 \subset \mathcal{H}$ is a strictly smaller subset, then it's possible that $h^* \in \mathcal{H}$, but $h^* \notin \mathcal{H}_0$.
- Moreover, it's likely that the best predictor in $\mathcal{H}_0$ (h_0^*, say) is worse than h^* (i.e., $R(h_0^*) > R(h^*)$), possibly much worse.
- So at the moment, it's not obvious why we would want to restrict $\mathcal{H}$, but it will become clear presently.
- Restrictions on $\mathcal{H}$ are sometimes called limitations on the "flexibility," "capacity, " or "roughness" of $h(\mathbf{x})$.

Best Constant Prediction with MSE Loss

- Suppose you have just drawn $(x, y) \in \mathbb{R} \times \mathbb{R}$ from a bivariate population, but you haven't looked at it yet.
- Your job is to make your best prediction $h = \widehat{y}$ (a scalar) of the value of y in the draw without using the information in x at all.
- That is, we are restricting the available hypotheses $\mathcal{H}$ to $\mathcal{H}_0$, the set of constant functions $h \in \mathbb{R}$.
- We adopt the MSE loss function, $\ell(\widehat{y}, y) = (\widehat{y} - y)^2$, which penalizes big errors much more than small ones.
- That choice results in risk $R(h) = \mathbb{E}_{x,y}[(h - y)^2]$, so the optimization problem which defines the Bayes optimal hypothesis relative to $\mathcal{H}_0$ is

$$\min_{h \in \mathcal{H}_0} \mathbb{E}_{x,y}[\ell(h, y)] = \min_{h \in \mathbb{R}} \mathbb{E}_y[(h - y)^2]$$

Best Constant Prediction with MSE Loss

- Under fairly general conditions, we can differentiate under the expectation.
- A differential characterization of the solution sets the first order conditions to zero, so $\mathbb{E}[2(h-y)]=0$, or $h^*=\mathbb{E}[y]\equiv\mu_y$.
- That is, under MSE loss, the best constant predictor of y is the population mean on y, μ_y.
- Under LAD loss, the best constant predictor is the median.

Best "Linear" Prediction with MSE Loss

- Now suppose I instead tell you that I want your best guess of y, but you are allowed to peek at the value of x when you make your prediction $h(x)$, but only to use the information in a limited way.
- That is, you are limited to $h \in \mathcal{H}_1 \subset \mathcal{H}$, the set of all affine functions of x.
- So $h(x) = \alpha + \beta x$, where the parameters α, β are to be determined.
- As before, we find the Bayes optimal hypothesis relative to $\mathcal{H}_1$ by minimizing risk:

$$\min_{h \in \mathcal{H}_1} \mathbb{E}_{x,y}[\ell(h(x), y)] = \min_{h \in \mathcal{H}_1} \mathbb{E}_{x,y}[(y - h(x))^2] = \min_{(\alpha,\beta) \in \mathbb{R}^2} \mathbb{E}_{x,y}[(y - \alpha - \beta x)^2]$$

Best "Linear" Prediction with MSE Loss

- Passing the derivative through the expectation, we have first order conditions

$$\mathbb{E}_{x,y}\left[\frac{\partial(y-\alpha-\beta x)^2}{\partial\alpha}\right] = \mathbb{E}_{x,y}[2(y-\alpha-\beta x)] = 0,$$

and

$$\mathbb{E}_{x,y}\left[\frac{\partial(y-\alpha-\beta x)^2}{\partial\beta}\right] = \mathbb{E}_{x,y}[2(y-\alpha-\beta x)x] = 0.$$

- So

$$\alpha = \mathbb{E}_y[y] - \beta\mathbb{E}_x[x] = \mu_y - \beta\mu_x$$

and

$$\beta = \mathbb{C}[x,y]/\mathbb{V}[x] = \frac{\sigma_{xy}}{\sigma_x^2} = \rho\frac{\sigma_y}{\sigma_x}$$

after plugging the first equation into the second and applying identities. Look familiar? Note that we are still talking about prediction in the population!

A "New Parameter" for Bivariate Populations

- We've seen that most univariate populations have a set of well-known parameters that help characterize them:
  - The population mean (μ).
  - The population variance (σ^2).
  - The population standard error (σ).
- Bivariate populations have additional interesting parameters, like the population correlation coefficient (ρ) we talked about earlier.
- The population linear projection or linear predictor (α, β) is just another parameter of a bivariate populations.

The Population Linear Projection

- The population linear projection for a bivariate population is a pair of numbers (α, β) giving the Bayes optimal predictor of y relative to the restricted hypothesis space $\mathcal{H}_1$ of affine functions:

$$\mathbb{L}[y|1, x] = \alpha + \beta x$$

- Note how similar the formula for the coefficients in the sample linear regression is to the formula for the population linear projection – that's the analogy principle in operation (more on that momentarily).
- BUT REMEMBER: $\alpha \neq \widehat{\alpha}$ and $\beta \neq \widehat{\beta}$! A parameter is different than an estimate!
- Also, unlike $\mathbb{E}$ for expectation, the use of $\mathbb{L}$ for the linear predictor is not standard across authors/texts.

The Population Linear Projection (with Matrices)

- When we have multiple regressors $\mathbf{x}_{1\times k}$ (a row vector), the problem setup is very similar.
- We are looking for a best predictor h^* in the restricted space of predictors $\mathcal{H}_1 \subset \mathcal{H}$ having $h(\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$ where $\boldsymbol{\beta}_{k\times 1}$ is a k-dimensional vector.
- Risk minimization over $\mathcal{H}_1$ gives

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^k} \mathbb{E}[(y - \mathbf{x}\boldsymbol{\beta})^2]$$

- Like the one-covariate case, we use a differential characterization of the solution.
- The minimum occurs where the differential in all k directions is zero – at the "bottom of the bowl."

The Population Linear Projection (with Matrices)

- Differentiating under the expectation, rules for vector derivatives imply first order conditions

$$\mathbb{E}_{\mathbf{x},y}\left[\frac{\partial(y-\mathbf{x}\boldsymbol{\beta})^2}{\partial\boldsymbol{\beta}}\right] = \mathbb{E}_{\mathbf{x},y}[2(y-\mathbf{x}\boldsymbol{\beta})\mathbf{x}] = \mathbf{0}_{1\times k}$$

- Taking a transpose gives $\mathbb{E}_{\mathbf{x},y}[\mathbf{x}^\top(y-\mathbf{x}\boldsymbol{\beta})] = \mathbf{0}_{k\times 1}$, and applying the rules for expectations gives

$$\mathbb{E}_{\mathbf{x},y}[\mathbf{x}^\top y] - \mathbb{E}_{\mathbf{x},y}[\mathbf{x}^\top(\mathbf{x}\boldsymbol{\beta})] = \mathbf{0}_{k\times 1}$$

- Using the associative rule and the fact that $\boldsymbol{\beta}$ is not random, we have

$$\boldsymbol{\beta} = (\mathbb{E}_{\mathbf{x},y}[\mathbf{x}^\top\mathbf{x}])^{-1}\mathbb{E}_{\mathbf{x},y}[\mathbf{x}^\top y].$$

Prediction: The Population Linear Projection (Matrices)

As an example, take $\mathbf{x} = \begin{bmatrix} 1 & x \end{bmatrix}$, so we are trying to predict y from the bivariate population (x, y). Then expanding $\mathbb{E}_{x,y}[\mathbf{x}^\top \mathbf{x}]$ gives

$$\mathbb{E}_x \left[\begin{bmatrix} 1 \\ x \end{bmatrix} \begin{bmatrix} 1 & x \end{bmatrix} \right] = \mathbb{E}_x \left[\begin{bmatrix} 1 & x \\ x & x^2 \end{bmatrix} \right] = \begin{bmatrix} \mathbb{E}_x[1] & \mathbb{E}_x[x] \\ \mathbb{E}_x[x] & \mathbb{E}_x[x^2] \end{bmatrix} = \begin{bmatrix} 1 & \mu_x \\ \mu_x & \sigma_x^2 + \mu_x^2 \end{bmatrix}$$

and similarly $\mathbb{E}_{x,y}[\mathbf{x}^\top y]$ gives

$$\mathbb{E}_{x,y} \left[\begin{bmatrix} 1 \\ x \end{bmatrix} y \right] = \mathbb{E}_{x,y} \left[\begin{bmatrix} y \\ xy \end{bmatrix} \right] = \begin{bmatrix} \mathbb{E}_{x,y}[y] \\ \mathbb{E}_{x,y}[xy] \end{bmatrix} = \begin{bmatrix} \mu_y \\ \sigma_{xy} + \mu_x \mu_y \end{bmatrix}$$

Prediction: The Population Linear Projection (Matrices)

Now taking

$$\boldsymbol{\beta} = (\mathbb{E}_{x,y}[\mathbf{x}^\top \mathbf{x}])^{-1} \mathbb{E}_{x,y}[\mathbf{x}^\top y] = \begin{bmatrix} 1 & \mu_x \\ \mu_x & \sigma_x^2 + \mu_x^2 \end{bmatrix}^{-1} \begin{bmatrix} \mu_y \\ \sigma_{xy} + \mu_x \mu_y \end{bmatrix}$$

and using our rules for matrix inverses and multiplication gives

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \mu_y - \frac{\sigma_{xy}}{\sigma_x^2} \mu_x \\ \frac{\sigma_{xy}}{\sigma_x^2} \end{bmatrix}$$

With the replacement of (α, β) by (β_1, β_2), this is precisely our result from earlier.

A Note on "Linear"ity

- In our derivation of the Bayes optimal hypothesis $\boldsymbol{\beta}$, we differentiated only with respect to $\boldsymbol{\beta}$, not $\mathbf{x}$.
- The derivation goes through as long as $\mathbb{L}[y|1, \mathbf{x}]$ is linear in $\boldsymbol{\beta}$.
- So, in particular, $\mathbf{x}$ could contain (many) nonlinear functions of any covariates $\{x_k\}$.
- x_k, x_k^2, x_k^r, $\log x_k$, $x_k x_j$, etc. are all fair game.
- "Linear" projection means linear in the parameters.

Best Prediction under MSE Loss

- Returning to our earlier problem, suppose finally that I just want your best prediction of y; you are now allowed to use the information x in any way at all in the hypothesis $h(\cdot)$.
- Since we already noted that a wide variety of nonlinear functions of covariates are permissible in a "linear" design, what new is being added here?
- Functions like $x_j^{\beta_j}$ or $\sin(\beta_j x_j)$, which are nonlinear in the parameters, are now permissible.
- Any h in $\mathcal{H}$ is now fair game.
- In principle, this space of possibly wild and misbehaved functions could be a problem if we want to guarantee a solution is concerned.[3] In practice, this turns out not to be a problem, as we will now see.

[3] If you are sufficiently mathematically sophisticated to understand why, good for you!

Best Prediction under MSE Loss

- Let $h \in \mathcal{H}$ denote any hypothesis and solve

$$\begin{aligned}\min_{h \in \mathcal{H}} \mathbb{E}_{x,y}[\ell(h(x), y)] &\equiv \min_{h \in \mathcal{H}} \mathbb{E}_{x,y}[(y - h(x))^2] \\ &= \mathbb{E}[\mathbb{E}[(y - h(x))^2 | x]] \\ &= \mathbb{E}[\mathbb{V}[y|x] + (\mathbb{E}[y - h(x)|x])^2].\end{aligned}$$

- The second equality is the Law of Iterated Expectations.
- The third equality uses the identity relating the variance to the second raw moment of a r.v.: $\mathbb{V}[Z] = \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2$.
- Evidently, $\mathbb{V}[y|x]$ does not depend on h at all, so it's clear that the expression takes a pointwise minimum when $\mathbb{E}[y - h(x)|x] = 0$; that is, where $h^*(x) = \mathbb{E}[y|x]$.
- The function $h^*(x)$ is called the regression function, or the conditional expectation function (CEF).

A Concrete Example

- Suppose that $F_{x,y}$ is bivariate normal $\mathcal{N}(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \sigma_{xy})$.
- We know that the CEF is linear, so

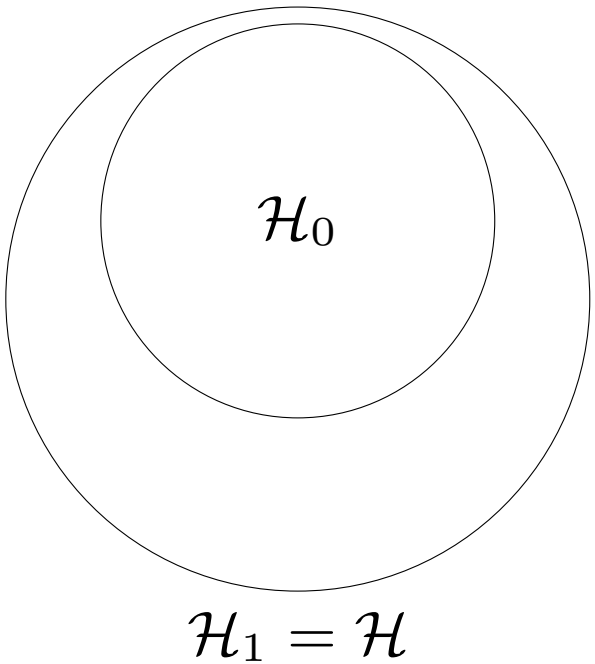
$$h^* = \mathbb{E}[y|x] = \mathbb{L}[y|x] = \beta_0 + \beta_1 x,$$

  with $\beta_0 \equiv \mu_y - \beta_1 \mu_x$ and $\beta_1 \equiv \frac{\sigma_{xy}}{\sigma_x^2}$.
- Since the minimum-risk hypothesis is linear, $h^* = h_1^*$, and $h^* \in \mathcal{H}_1$.
- But unless $\beta_1 = 0$, the best constant predictor $h_0^* = \mu_y \neq h_1^*$, so $h^* \notin \mathcal{H}_0$.
- Therefore, if we use $h_0^* = \mu_y$ rather than $h_1^* = h^* = \beta_0 + \beta_1 x$, as our predictor, we incur an approximation error equal to

$$h_0 - h_1 = \mu_y - (\beta_0 + \beta_1 x) = \mu_y - (\mu_y - \beta_1 \mu_x) - \beta_1 x = \beta_1(\mu_x - x).$$

- This is the error from using the wrong predictor. It is reducible in the sense that we could do better than $\mathcal{H}_0$ by using $\mathcal{H}_1$ instead. But this is still a population exercise (no estimation yet).

Relationship Between $\mathcal{H}$, $\mathcal{H}_1$, and $\mathcal{H}_0$



Best Prediction under 0-1 Loss

- Now consider the complementary problem where $\mathcal{X} = \mathbb{R}$ again, but $\mathcal{Y} = \{0, 1\}$ and you are to select a hypothesis $h : \mathcal{X} \to \mathcal{Y}$ which is a best predictor of y given the information in x.
- Since y is 0 or 1, so is h.
- This problem is called classification, and should be familiar if you think back to your work with logit or probit models.
- To simplify notation, we define the probability of outcome $y = 1$ given x as $\eta(x) \equiv F_{y|x}(y = 1)$.
- We will now characterize the Bayes optimal predictor under 0-1 loss, sometimes called the Bayes classifier, or oracle predictor.

Best Prediction under 0-1 Loss

- Let $h \in \mathcal{H}$ denote any hypothesis and solve

$$\begin{aligned}
\min_{h\in\mathcal{H}} R(h) &= \min_{h\in\mathcal{H}} \mathbb{E}_{x,y}[\ell(h(x), y)] \\
&= \min_{h\in\mathcal{H}} \mathbb{E}_{x,y}[\mathbb{1}[y \neq h(x)]] \\
&= \mathbb{E}[\mathbb{E}[\mathbb{1}[y \neq h(x)]|x]] \\
&= \mathbb{E}[\mathbb{E}[\mathbb{1}[y = 1, h(x) = 0] + \mathbb{1}[y = 0, h(x) = 1]|x]] \\
&= \mathbb{E}[\mathbb{E}[\mathbb{1}[y = 1]\mathbb{1}[h(x) = 0] + \mathbb{1}[y = 0]\mathbb{1}[h(x) = 1]|x]] \\
&= \mathbb{E}[\eta(x)\mathbb{1}[h(x) = 0] + (1 - \eta(x))\mathbb{1}[h(x) = 1]] \\
&= \mathbb{E}[\eta(x)(1 - \mathbb{1}[h(x) = 1]) + (1 - \eta(x))\mathbb{1}[h(x) = 1]] \\
&= \mathbb{E}[\mathbb{1}[h(x) = 1](1 - \eta(x) - \eta(x)) + \eta(x)] \\
&= \mathbb{E}[\mathbb{1}[h(x) = 1](1 - 2\eta(x)) + \eta(x)],
\end{aligned}$$

- Since $\eta(x)$ does not depend on h, the expression is minimized (again, pointwise in x), by having $h(x) = 1$ only when $1 - 2\eta(x) < 0$ (so $h(x) = 0$ when $1 - 2\eta(x) > 0$).
- That is, we predict $h(x) = 1$ whenever $\eta(x) > \frac{1}{2}$, or $F_{y|x}(y = 1) > \frac{1}{2}$.

One More Loss Function: The Kullback-Leibler (KL) Divergence

- If y is not uniquely determined by x , i.e. if there does not exist a deterministic function h such that $y = h(x)$, then the relation between x and y needs to be described by a joint-distribution $p(x, y)$.
- Since $p(x)$ is observable, we often just look for $p(y|x)$, since $p(x, y) = p(y|x)p(x)$.
- To learn $p(y|x)$, let the hypothesis set $\mathcal{H}$ be a set of conditional probability distributions: $\mathcal{H} = \{q_1(y|x), q_2(y|x), \ldots\}$.
- Our goal is to select a $q(y|x)$ from $\mathcal{H}$ that approximates the true $p(y|x)$ tolerably well. How can we quantify how well $q(y|x)$ approximates $p(y|x)$?

A Gentle Introduction to Information Theory

- How much information is conveyed when we observe a specific value of a random variable?
- A highly improbable outcome conveys more information when it happens than a very common one. Information is related to "surprise."
- If we know an event is certain to happen, we would receive no information when we observe it.
- Something to think about next time you drop your "outliers."

A Gentle Introduction to Information Theory

Let $H(\cdot)$ denote the information content of an event. $H(\cdot)$ should satisfy a couple of axioms:

- $H(A)$ should be inversely correlated with the probability of the event $p(A)$.
- For two unrelated events A and B, with $p(AB) = p(A)p(B)$, the information on the joint event should be strictly additive: $H(AB) = H(A) + H(B)$.

One simple function which fits both criteria is

$$H(A) = \log\left(\frac{1}{p(A)}\right) = -\log p(A)$$

Entropy

For a discrete random variable with probability distribution $p(y)$, the average amount of information transmitted is

$$\mathbb{H}(p) = \mathbb{E}_p[H] = \sum_y -\log(p(y))p(y)$$

The quantity $\mathbb{H}(p)$ is called the entropy of the distribution function $p(\cdot)$.

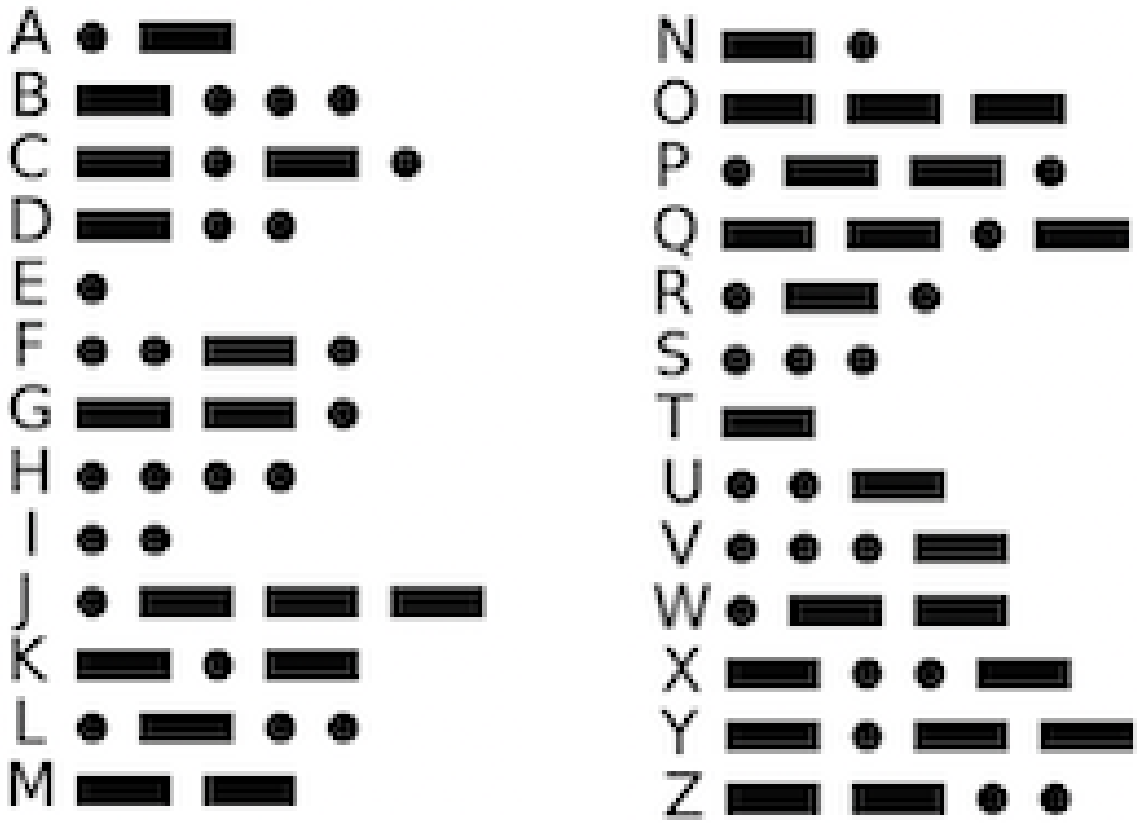
Distributions that have sharp peaks (are highly concave) around some values will have relatively low entropy, while those that are less concave (more spread out) have higher entropy.

Entropy

- Historically, information entropy was developed to describe the average amount of information needed to specify the state of a random variable.
- Specifically, if we use base-2 logs in the definition of $\mathbb{H}(p)$, then $\mathbb{H}(p)$ is a lower bound on the average number of bits needed to encode a random variable with probability distribution p.
- Achieving this bound would require using an optimal coding scheme designed for p, which assigns shorter codes to higher probability events and longer codes to less probable events.

Morse Code

What's the most common letter in the English language? The least?



Entropy Example

- Suppose a random variable has 4 possible states, with each state being equally likely. Then we can code these 4 states as

$$00 \quad 01 \quad 10 \quad 11$$

  The average number of bits needed to encode the state is 2, which is equal to the entropy $\mathbb{H} = -\sum_s \log_2(p_s) \cdot p_s = 4 \times (2 \times \frac{1}{4}) = 2$.

- If the probabilities of the 4 states change, say, to $(0.4, 0.35, 0.2, 0.05)$, then the entropy changes to $\mathbb{H} = -\sum_s \log_2(p_s) \cdot p_s = 1.739$.

- A better coding scheme (e.g., Huffman coding) can help us do a little better. Taking the alphabet to be "0","10","110","111" gives an average length of $.4(1) + .35(2) + .2(3) + .05(3) = 1.85$.

Cross Entropy

Now suppose p is the true distribution of our random variable, but we instead use element q from our hypothesis space. The average information needed to encode p using the "wrong" distribution q is given by

$$\mathbb{H}(p, q) = \mathbb{E}_p[-\log q] = \sum_y -\log(q(y))p(y)$$

$\mathbb{H}(p, q)$ is called the cross entropy of p and q.

Kullback-Leibler Divergence

The relative entropy of q with respect to p, also known as the Kullback-Leibler (KL) divergence of q from p is defined as

$$D_{KL}(p||q) = \mathbb{H}(p,q) - \mathbb{H}(p)$$
$$\sum_y \log\left(\frac{p(y)}{q(y)}\right) p(y)$$

in the case of discrete y, or for continuous y as

$$D_{KL}(p||q) = \int \log\left(\frac{p(y)}{q(y)}\right) p(y) dy$$

The KL divergence represents the average additional information required to express the random variable y as a result of using q rather than the true distribution p.

Example, Continued

Continuing the earlier example, let the true distribution $p = (0.4, 0.35, 0.2, .05)$ and the hypothetical distribution be $q = (.25, .25, .25, .25)$.

- We already know that the entropy of the true distribution is 1.739 ("bits"). In $\log_e$ units ("nits") it is 1.206.
- The cross entropy of using q for an optimal p is

$$\begin{aligned}\mathbb{H}(p, q) &= .4 \cdot (-\log(.25)) + .35 \cdot (-\log(.25)) + \\ &\qquad .2 \cdot (-\log(.25)) + .05 \cdot (-\log(.25)) \\ &= 1.38\end{aligned}$$

- So the KL divergence is given by

$$D_{KL}(p||q) = \mathbb{H}(p, q) - \mathbb{H}(p) = 1.386 - 1.206 = 0.18$$

KL Divergence as a Loss Function

- Consequently, the KL Divergence can be interpreted as a measure of the dissimilarity between a distribution from our hypothesis space, $q \in \mathcal{H}$ and the true distribution p.
- It satisfies $D_{KL}(p||q) \geq 0$ iff $p = q$.
- But it is not a distance since it doesn't satisfy the triangle inequality.
- Even so, it can be used as a loss function to quantify how much q diverges from p, which is what we wanted.

Prediction: Conclusions

- The loss function $\ell(\cdot, \cdot)$ you employ is your choice – and it's important: it defines a poor prediction and the consequences.
- There are many loss functions (including asymmetric ones) we have not covered here, so it's worth thinking about the circumstances of your own particular problem before blindly applying MSE, even though MSE is often a very workable choice.
- The hypothesis space $\mathcal{H}$ is also your choice, but currently we don't have any incentive to make it smaller than "everything."
- Once $\ell(\cdot, \cdot)$ and $\mathcal{H}$ are decided, the best predictor h^* (the "oracle") for $F_{x,y}$ can be identified.
- This best predictor is a natural target for estimation.
- Even though there may be a diversity of views as to the preferred estimation approach (as we will see), there's often broad agreement on the right target, and that's sometimes helpful to know.