Otsu's Binarization

Stephen Huan[1]

[1]Thomas Jefferson High School for Science and Technology

TJ Vision & Graphics Club, April 14, 2021

Probability Theory

Before we start, we'll need to know some basic probability...

Table of Contents

Basic Definitions

Definition

The *sample space* is the set of all possible outcomes, commonly denoted Ω. An *event* is just "something which occurs", or formally speaking, a set which is a subset of Ω.

Definition

The *class of events* $\mathcal{F}$ is a σ-algebra on Ω (that is, it is a collection of the subsets of Ω, including Ω, and closed under union and complement). We will assume $\mathcal{F} = \mathcal{P}(\Omega)$, the power set of Ω (the set of all subsets of Ω).

Probability Function

Definition

(Kolmogorov axioms) Given a sample space Ω and event class $\mathcal{F}$, a probability function P has the following properties:

1. $P(E) \in \mathbb{R}, P(E) \geq 0$ for all $E \in \mathcal{F}$
2. $P(\Omega) = 1$
3. If E_1 and E_2 are disjoint sets in $\mathcal{F}$, $P(E_1 \cup E_2) = P(E_1) + P(E_2)$

It follows that $(\Omega, \mathcal{F}, P)$ is a *probability space*.

Law of Total Probability

Theorem

If B_i is a partition of Ω, then $P(A) = \sum_i P(A \cap B_i)$

Proof.

Let's look at the sum of the first two terms.
$P(A \cap B_1) + P(A \cap B_2) = P((A \cap B_1) \cup (A \cap B_2))$ because $A \cap B_1$ and $A \cap B_2$ are disjoint (B_1 and B_2 are disjoint).
$= P(A \cap (B_1 \cup B_2))$. Extending the logic to all B_i,

$$= P(A \cap (B_1 \cup B_2 \cup \cdots \cup B_n))$$

We know $B_1 \cup B_2 \cup \cdots \cup B_n = \Omega$ since it's a partition, and $A \cap \Omega = A$ by definition. So this is just $P(A)$. $\square$

Table of Contents

Random Variable

Definition

A *random variable* (r.v.) is a function $\Omega \to \mathbb{R}$, i.e. a function assigning a number to each outcome, but can more intuitively be thought of as a "dispenser" of values. We will denote random variables as a single uppercase letter, e.g. X or Z.

Randomness

Although the probability space carries "randomness", a random variable is neither random nor a variable — it is a deterministic function assigning a fixed number to a particular outcome. Although what outcome you get is random (from the randomness of the space), the assignment is the same.

Expected Value

Definition

The *expected value* is what we "expect" a random variable to dispense over many samples, or the average value:

$$\mathrm{E}[X] = \sum_{x \in X} x\, P(x)$$

We will also denoted the expected value as μ.

Example

Let X take on value 1 with probability $\frac{1}{2}$, 2 with $\frac{1}{4}$ chance, and 3 with $\frac{1}{4}$ chance. $\mathrm{E}[X] = 2 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{4} = \frac{9}{4}$.

Law of the Unconscious Statistician

Theorem

$\mathrm{E}[f(X)] = \sum_{x \in X} f(x)\, P(x)$, *i.e. the expected value of a transformation of a random variable.*

Proof.

(informal) We can think of $f(x)$ as partitioning X into subsets like $S = \{x \in X \mid f(x) = y\}$ for some y. When we compute $P(y)$, it'll be the sum of $P(x)$ for all $x \in S$. But we know each $S(y)$ is disjoint for different values of y since they necessarily form a partition. So if we iterate over x, we cover the same set anyways, multiplying each $P(x)$ by the $f(x)$ it belongs to. $\square$

Linearity of Expectation

Corollary

The linearity of expectation, i.e.

1. $\mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y]$ *for random variables* X, Y
2. $\mathrm{E}[cX] = c\,\mathrm{E}[X]$ *for* $c \in \mathbb{R}$

Proof of Linearity of Expectation

Proof.

Begin with the definition of expectation:

$$\begin{aligned}
\mathrm{E}[X+Y] &= \sum_{x\in X}\sum_{y\in Y}(x+y)\,P(x,y) \\
&= \sum_{x\in X}\sum_{y\in Y} x\,P(x,y) + \sum_{x\in X}\sum_{y\in Y} y\,P(x,y)
\end{aligned}$$

Swapping order and applying the law of total probabilities,

$$\begin{aligned}
&= \sum_{x\in X} x \underbrace{\sum_{y\in Y} P(x,y)}_{P(x)} + \sum_{y\in Y} y \underbrace{\sum_{x\in X} P(x,y)}_{P(y)} \\
&= \sum_{x\in X} x\,P(x) + \sum_{y\in Y} y\,P(y) = \mathrm{E}[X] + \mathrm{E}[Y]
\end{aligned}$$

Proof of the Linearity of Expectation, Continued

Proof.

We use the transformation $f(x) = cx$ with the law of the unconscious statistician:

$$\begin{aligned} \mathrm{E}[cX] &= \sum_{x \in X} (cx)\, P(x) \\ &= c \sum_{x \in X} x\, P(x) \\ &= c\, \mathrm{E}[x] \end{aligned}$$

$\square$

Variance

Definition

The *variance* is the expected squared deviation from the expected value. The larger the variance, the more "variable" the random variable is. By definition,

$$\mathrm{Var}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2]$$

We also denote the variance as σ^2, since it is the square of the standard deviation σ.

Useful Alternative Form

Theorem

$\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$, *a convenient form of variance.*

Proof.

By definition,

$$\begin{aligned}\mathrm{Var}[X] &= \mathrm{E}[(X - \mathrm{E}[X])^2] \\ &= \mathrm{E}[X^2 - 2X\,\mathrm{E}[X] + \mathrm{E}[X]^2]\end{aligned}$$

From the linearity of expected value,

$$\begin{aligned}&= \mathrm{E}[X^2] - \mathrm{E}[2\,\mathrm{E}[X]X] + \mathrm{E}[\mathrm{E}[X]^2] \\ &= \mathrm{E}[X^2] - (2\,\mathrm{E}[X])\,\mathrm{E}[X] + \mathrm{E}[X]^2 \\ &= \mathrm{E}[X^2] - \mathrm{E}[X]^2\end{aligned}$$

$\square$

Table of Contents

k-means Refresher

Given a set of points X, partition into S_i. Cost function:

$$\sum_i \sum_{\boldsymbol{x} \in S_i} \|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2$$

Recall the definition of

$$\mathrm{Var}[S_i] = \mathrm{E}[(S_i - \boldsymbol{\mu}_i)^2] = \sum_{\boldsymbol{x} \in S_i} P(\boldsymbol{x}) \|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2$$

Looks awfully similar, but we're missing a $P(\boldsymbol{x}) = \frac{1}{|S_i|}$:

$$\sum_i |S_i| \,\mathrm{Var}[S_i]$$

Dividing by the total number of points $|\mathrm{X}|$,

$$\sum_i \frac{|S_i|}{|X|} \mathrm{Var}[S_i] = \sum_i P(S_i) \,\mathrm{Var}[S_i]$$

$$= \boxed{\mathrm{E}[\mathrm{Var}[X \mid S]]}$$

Conditional Probability

Definition

The probability that an event A happens *conditional* on B is denoted $P(A \mid B)$, i.e. the probability A happens "given" B happens. This should not be confused for $P(A \cap B)$, the probability A happens *and* B happens.

Conditional Probability

Corollary

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Proof.

(informal) We will use an informal definition of probability as the number of ways for something to occur over the total number of ways. Since we know B occurs, the total number of ways is $P(B)$. The number of ways A occurs is the number of events where A occurs *and* B occurs, because we know B occurs. $\square$

Conditional Expectation

Definition

$\mathrm{E}[X \mid Y]$ is the *conditional expectation* of the random variable X conditional on the random variable Y. Note that $\mathrm{E}[X \mid Y]$ is a random variable which depends on the particular value of Y! Following a similar definition to expectation,

$$\mathrm{E}[X \mid Y] = \sum_{x \in X} x\, P(X = x \mid Y = y)$$

Example of conditional expectation

For example, $\mathrm{E}[X \mid Y = 1]$ is a real number equal to the expected value of X when Y is 1. But Y can take on many different values, so $\mathrm{E}[X \mid Y]$ in general is a random variable; by definition it assigns a real number to each outcome of Y.

Law of Total Expectation

Because $\mathrm{E}[X \mid Y]$ is a random variable, we can treat it like any other random variable and take its expectation and variance.

Theorem

$\mathrm{E}[X] = \mathrm{E}[\mathrm{E}[X \mid Y]]$ *for any random variables* X, Y

Proof of the Law of Total Expectation

Proof.

Starting with the definition of expectation,

$$\mathrm{E}[\mathrm{E}[X \mid Y]] = \sum_{y \in Y} P(y) \, \mathrm{E}[X \mid Y]$$

Expanding with the defintition of conditional expectation,

$$\begin{aligned}
&= \sum_{y \in Y} P(y) \sum_{x \in X} x \, P(x \mid y) \\
&= \sum_{y \in Y} \sum_{x \in X} x \, P(x, y) \\
&= \sum_{x \in X} x \sum_{y \in Y} P(x, y) \\
&= \sum_{x \in X} x \, P(x) = \mathrm{E}[X]
\end{aligned}$$

$\square$

Conditional Variance

Definition

$\text{Var}[X \mid Y]$ is the *conditional variance* of the random variable X conditional on the random variable Y. Like the conditional expectation, this is a random variable:

$$\text{Var}[X \mid Y] = \text{E}[(X - \text{E}[X \mid Y])^2 \mid Y]$$

Note: normally, $\text{Var}[X] = \text{E}[(X - \text{E}[X])^2]$. But we can't literally plug in $X \mid Y$ into that formula because $X \mid Y$ has no definition by itself: it must be wrapped in E or Var or some other operator to become a valid random variable.

Conditional Variance Alternative Form

Theorem

$$\operatorname{Var}[X \mid Y] = \operatorname{E}[X^2 \mid Y] - \operatorname{E}[X \mid Y]^2$$

Proof.

Begin with the definition of conditional variance,

$$\begin{aligned}
\operatorname{Var}[X \mid Y] &= \operatorname{E}[(X - \operatorname{E}[X \mid Y])^2 \mid Y] \\
&= \operatorname{E}[X^2 - 2X \operatorname{E}[X \mid Y] + \operatorname{E}[X \mid Y]^2 \mid Y] \\
&= \operatorname{E}[X^2 \mid Y] - 2 \operatorname{E}[X \mid Y] \operatorname{E}[X \mid Y] + \operatorname{E}[X \mid Y]^2 \\
&= \operatorname{E}[X^2 \mid Y] - \operatorname{E}[X \mid Y]^2
\end{aligned}$$

$\square$

Let's Play Around With This

What if we take the expectation of the conditional variance?

$$\begin{aligned}
\mathrm{E}[\mathrm{Var}[X \mid Y]] &= \mathrm{E}[\mathrm{E}[X^2 \mid Y] - \mathrm{E}[X \mid Y]^2] \\
&= \mathrm{E}[\mathrm{E}[X^2 \mid Y]] - \mathrm{E}[\mathrm{E}[X \mid Y]^2]
\end{aligned}$$

Using the law of total expectation,

$$= \mathrm{E}[X^2] - \mathrm{E}[\mathrm{E}[X \mid Y]^2]$$

Re-arranging,

$$\begin{aligned}
\mathrm{E}[X^2] &= \mathrm{E}[\mathrm{Var}[X \mid Y]] + \mathrm{E}[\mathrm{E}[X \mid Y]^2] \\
&= \mathrm{E}[\mathrm{Var}[X \mid Y] + \mathrm{E}[X \mid Y]^2]
\end{aligned}$$

Let's See Where This Goes

Obviously,

$$\mathrm{E}[X]^2 = \mathrm{E}[\mathrm{E}[X \mid Y]]^2$$

So we have an alternative form for $\mathrm{E}[X^2]$ and $\mathrm{E}[X]^2$ in terms of conditionals. What if we put them together?

$$\begin{aligned}\mathrm{Var}[X] &= \mathrm{E}[X^2] - \mathrm{E}[X]^2 \\ &= \mathrm{E}[\mathrm{Var}[X \mid Y] + \mathrm{E}[X \mid Y]^2] - \mathrm{E}[\mathrm{E}[X \mid Y]]^2 \\ &= \mathrm{E}[\mathrm{Var}[X \mid Y]] + (\mathrm{E}[\mathrm{E}[X \mid Y]^2] - \mathrm{E}[\mathrm{E}[X \mid Y]]^2)\end{aligned}$$

But $\mathrm{E}[X \mid Y]$ is a normal random variable, so

$$\mathrm{Var}[X] = \boxed{\mathrm{E}[\mathrm{Var}[X \mid Y]] + \mathrm{Var}[\mathrm{E}[X \mid Y]]}$$

This is the law of total variance!

Law of Total Variance

Theorem

$$\mathrm{Var}[X] = \underbrace{\mathrm{E}[\mathrm{Var}[X \mid Y]]}_{\textit{intra-class variance}} + \underbrace{\mathrm{Var}[\mathrm{E}[X \mid Y]]}_{\textit{inter-class variance}}$$

Discussion

$\mathrm{E}[\mathrm{Var}[X \mid Y]]$ measures *intra*-class variance because it's the expected variance of each group (the lower the variance of each group, the lower the expected value). Likewise, $\mathrm{Var}[\mathrm{E}[X \mid Y]]$ measures *inter*-class variance. The more different the groups are from each other (measured by the difference between their means), the larger the variance.

Back to k-means

Recall that we got k-means's cost function into the form $\mathrm{E}[\mathrm{Var}[X \mid S]]$. Let's see what that means.

Back to k-means

Recall that we got k-means's cost function into the form $\mathrm{E}[\mathrm{Var}[X \mid S]]$. Let's see what that means.
Define S to be a random variable (a function from outcomes to real numbers) that assigns integer values to each point $\boldsymbol{x} \in X$, such that two points have the same $S(\boldsymbol{x})$ if and only if they're in the same partition.

Back to k-means

Recall that we got k-means's cost function into the form $\mathrm{E}[\mathrm{Var}[X \mid S]]$. Let's see what that means.
Define S to be a random variable (a function from outcomes to real numbers) that assigns integer values to each point $\boldsymbol{x} \in X$, such that two points have the same $S(\boldsymbol{x})$ if and only if they're in the same partition.
It may not seem morally correct to have S be a random variable when it has no "randomness" whatsoever. Well, we converted all the sums to expectation and variance already, so we can just fit it into the tools we have.

k-means Analysis

$$\mathrm{E}[\mathrm{Var}[X \mid S]] = \sum_i P(S = i)\,\mathrm{Var}[X \mid S = i]$$

where i is the values S can take on

$$= \sum_i \frac{|S_i|}{|X|} \sum_{\boldsymbol{x} \mid S(\boldsymbol{x}) = i} P(\boldsymbol{x}) \|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2$$

$$= \frac{1}{|X|} \sum_i |S_i| \sum_{\boldsymbol{x} \mid S(\boldsymbol{x}) = i} \frac{1}{|S_i|} \|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2$$

$$= \frac{1}{|X|} \sum_i \sum_{\boldsymbol{x} \mid S(\boldsymbol{x}) = i} \|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2$$

which is just the cost function of k-means, as expected

k-means interpretation

k-mean's cost function is $\mathrm{E}[\mathrm{Var}[X \mid S]]$, or the inter-class variance (we want to make the points inside each group as close to each other as possible). But we know from the law of total variance that $\mathrm{Var}[X] = \mathrm{E}[\mathrm{Var}[X \mid S]] + \mathrm{Var}[\mathrm{E}[X \mid S]]$

k-means interpretation

k-mean's cost function is $\mathrm{E}[\mathrm{Var}[X \mid S]]$, or the inter-class variance (we want to make the points inside each group as close to each other as possible). But we know from the law of total variance that $\mathrm{Var}[X] = \mathrm{E}[\mathrm{Var}[X \mid S]] + \mathrm{Var}[\mathrm{E}[X \mid S]]$
So *minimizing inter*-class variance is equivalent to *maximizing intra*-class variance. Making each group as similar as possible must make the groups different from each other.
There's only so much variance to go around!

Table of Contents

Binarization

- Suppose we have an image
- We can greyscale by averaging the channels
- Then binarize the image (make each pixel on or off)
- This is implicitly done in edge detection

Binary image

Figure: An image binarized with Otsu's binarization.

Cost Function

How should we measure the effectiveness?

Cost Function

How should we measure the effectiveness?
Well, we're splitting the pixels into two groups. Minimize the variance of each group, weighted by the size of the group.

Cost Function

How should we measure the effectiveness?
Well, we're splitting the pixels into two groups. Minimize the variance of each group, weighted by the size of the group. This is identical to k-means with $k = 2$.

Cost Function

How should we measure the effectiveness?
Well, we're splitting the pixels into two groups. Minimize the variance of each group, weighted by the size of the group.
This is identical to k-means with $k = 2$.
However, the standard k-means algorithm is suboptimal and may take longer than we want. Can we do optimal k-means efficiently if we know the data is one-dimensional and $k = 2$?

Otsu's Binarization, Justification

There are efficient algorithms for 1D k-means, see *Fast Exact k-Means, k-Medians and Bregman Divergence Clustering in 1D*.

Otsu's Binarization, Justification

There are efficient algorithms for 1D k-means, see *Fast Exact k-Means, k-Medians and Bregman Divergence Clustering in 1D*.
However, these algorithms are pretty complicated.
Can we make use of $k = 2$?

Otsu's Binarization, Justification

There are efficient algorithms for 1D k-means, see *Fast Exact k-Means, k-Medians and Bregman Divergence Clustering in 1D*.
However, these algorithms are pretty complicated.
Can we make use of $k = 2$?
We will never make a non-contiguous group, since if a group is split we might as well assign those points to the other group.
Another, more precise observation, is that if we place two centers, there's only one point equidistant to those centers, the midpoint of the centers. Anything to the left of the midpoint is assigned to one center and anything to the right the other.

Otsu's Binarization, Justification

There are efficient algorithms for 1D k-means, see *Fast Exact k-Means, k-Medians and Bregman Divergence Clustering in 1D*.
However, these algorithms are pretty complicated.
Can we make use of $k = 2$?
We will never make a non-contiguous group, since if a group is split we might as well assign those points to the other group.
Another, more precise observation, is that if we place two centers, there's only one point equidistant to those centers, the midpoint of the centers. Anything to the left of the midpoint is assigned to one center and anything to the right the other.
In 1D and with $k = 2$, we have exactly one *threshold* value.
How many possible thresholds are there?

Otsu's Binarization, Summary

- Iterate over the 256 possible thresholds
- Maintain statistics counters
- Compute the expected variance after each split
- Pick the threshold with the smallest intra-class variance

Table of Contents

Inter-class Variance

Standard approach: minimizing intra-class variance is maximizing inter-class variance.

By definition of variance,

$$\mathrm{Var}[\mathrm{E}[X \mid Y]] = \mathrm{E}[(\mathrm{E}[X \mid Y] - \mathrm{E}[\mathrm{E}[X \mid Y]])^2]$$

Using the law of total expectation,

$$= \mathrm{E}[(\mathrm{E}[X \mid Y] - \mathrm{E}[X])^2]$$

$$= \sum_i P(Y = i)(\mathrm{E}[X \mid Y = i] - \mathrm{E}[X])^2$$

To make the notation more concise, let the inter-class variance be σ_B^2, $P(Y = i)$ be ω_i, and $\mathrm{E}[X \mid Y = i]$ be μ_i:

$$\sigma_B^2 = \omega_0(\mu_0 - \mu)^2 + \omega_1(\mu_1 - \mu)^2$$

Inter-class Variance, Continued

$$\begin{aligned}\sigma_B^2 &= \omega_0(\mu_0 - \mu)^2 + \omega_1(\mu_1 - \mu)^2 \\ &= \omega_0[\mu_0^2 - 2\mu_0\mu + \mu^2] + \omega_1[\mu_1^2 - 2\mu_1\mu + \mu^2] \\ &= \omega_0\mu_0^2 + \omega_1\mu_1^2 - 2\mu(\omega_0\mu_0 + \omega_1\mu_1) + \mu^2(\omega_0 + \omega_1)\end{aligned}$$

We know $\omega_0 + \omega_1 = 1$ because they're probabilities and $\omega_0\mu_0 + \omega_1\mu_1 = \mathrm{E}[\mathrm{E}[X \mid Y]] = \mathrm{E}[X] = \mu$ so

$$\begin{aligned} &= \omega_0\mu_0^2 + \omega_1\mu_1^2 - \mu^2 \\ &= \omega_0\mu_0^2 + \omega_1\mu_1^2 - (\omega_0\mu_0 + \omega_1\mu_1)^2 \\ &= (\omega_0 - \omega_0^2)\mu_0^2 - 2\omega_0\omega_1\mu_0\mu_1 + (\omega_1 - \omega_1^2)\mu_1^2 \\ &= \omega_0(1 - \omega_0)\mu_0^2 - 2\omega_0\omega_1\mu_0\mu_1 + \omega_1(1 - \omega_1)\mu_1^2\end{aligned}$$

We know $1 - \omega_0 = \omega_1$ and likewise for $1 - \omega_1$,

$$= \omega_0\omega_1\mu_0^2 - 2\omega_0\omega_1\mu_0\mu_1 + \omega_0\omega_1\mu_1^2 = \boxed{\omega_0\omega_1(\mu_0 - \mu_1)^2}$$

Algorithm Overview

1. Precompute the number of pixels of each intensity
2. Iterate over possible thresholds, maintaining counters
3. Compute threshold which maximizes inter-class variance
4. Binarize image with optimal threshold

Image distribution

frequency counts of each pixel

```
def histogram(img: np.array) -> np.array:
    """ Computes the distribution of the image. """
    p = np.zeros(256, dtype=np.int)
    for x in img.flatten():
        p[x] += 1
    return p
```

Algorithm Details

- We need to keep track of ω_0 and μ_0
- Can compute ω_1 and μ_1 by subtraction
- Easier if they are not fractions but integer counts, i.e. $W_0 = |S_0| = |X|\omega_0$ and $A_0 = \sum_{x|S(x)=0} x = |S_0|\mu_0$
- We know we're scanning in terms of increasing threshold t
- So the only change is going to be pixels with intensity t
- Then the update is just $W_0 + p[t]$ and $A_0 + t\, p[t]$
- $\omega_0 = \frac{W_0}{|X|}$ and $\mu_0 = \frac{A_0}{W_0}$

Algorithm Otsu's binarization with inter-class variance

```
def otsu(img: np.array) -> np.array:
    """ Applies Otsu's binarization to the image. """
    h, n = histogram(img)
    X0, X1, p = 0, img.sum(), 0
    threshold, best = -1, 0
    for t in range(256):
        u1, u0 = t*h[t], h[t]
        X0, X1, p = X0 + u1, X1 - u1, p + u0
        if p > 0 and n - p > 0:
            # divide by n^2 to get the true variance
            var = p*(n - p)*(X0/p - X1/(n - p))**2
            if var > best:
                threshold, best = t, var
    return __threshold(img, threshold)
```

Binarization

Otsu's binarization computes the optimal threshold.
We want a binary image.

thresholding

```
def __threshold(img: np.array, t: int) -> np.array:
    """ Binarizes an image given a threshold. """
    return 255*(img > t)
```

Whether it's $>$ or $\geq$ is arbitrary, just needs to be consistent.

Table of Contents

Intra-class Variance

We just spent a lot of time deriving inter-class variance.
We needed to use the law of total variance to show that it's the same as minimizing intra-class variance.
Why not just directly attack intra-class variance?

By the definition of expectation,

$$\mathrm{E}[\mathrm{Var}[X \mid Y]] = \sum_i P(Y = i)\,\mathrm{Var}[X \mid Y = i]$$

As usual, let the intra-class variance be σ_W^2, $P(Y = i)$ be ω_i, and $\mathrm{Var}[X \mid Y = i]$ be σ_i^2:

$$\sigma_W^2 = \omega_0\sigma_0^2 + \omega_1\sigma_1^2$$

Intra-class Variance, Continued

Now let's re-write the conditional variances using $\mathrm{Var}[X \mid Y] = \mathrm{E}[X^2 \mid Y] - \mathrm{E}[X \mid Y]^2$:

$$\begin{aligned}\sigma_W^2 &= \omega_0\sigma_0^2 + \omega_1\sigma_1^2 \\ &= \omega_0[\frac{\sum_{x|S(x)=0} x^2}{|S_0|} - (\frac{\sum_{x|S(x)=0} x}{|S_0|})^2] \\ &+ \omega_1[\frac{\sum_{x|S(x)=1} x^2}{|S_1|} - (\frac{\sum_{x|S(x)=1} x}{|S_1|})^2]\end{aligned}$$

To make the notation more concise, let $W_i = |S_i|$ and $\sum x_i^k = \sum_{x|S(x)=i} x^k$:

$$= \frac{W_0}{|X|}[\frac{\sum x_0^2}{W_0} - (\frac{\sum x_0}{W_0})^2] + \frac{W_1}{|X|}[\frac{\sum x_1^2}{W_1} - (\frac{\sum x_1}{W_1})^2]$$

Intra-class Variance, Continued

$$\sigma_W^2 = \frac{W_0}{|X|}[\frac{\sum x_0^2}{W_0} - (\frac{\sum x_0}{W_0})^2] + \frac{W_1}{|X|}[\frac{\sum x_1^2}{W_1} - (\frac{\sum x_1}{W_1})^2]$$

$$= \frac{1}{|X|}([\sum x_0^2 - \frac{(\sum x_0)^2}{W_0}] + [\sum x_1^2 - \frac{(\sum x_1)^2}{W_1}])$$

Multiplying by $|X|$ and re-arranging,

$$= [\sum x_0^2 + \sum x_1^2] - [\frac{(\sum x_0)^2}{W_0} + \frac{(\sum x_1)^2}{W_1}]$$

The first term is just the sum of squares for each $x \in X$, which is a constant, and we can negate the second term, swapping the objective. Finally, putting in terms of the counter variables:

$$= \frac{(\sum x_0)^2}{W_0} + \frac{(\sum x_1)^2}{W_1} = \boxed{\frac{A_0^2}{W_0} + \frac{A_1^2}{|X| - W_0}}$$

Algorithm Otsu's binarization with intra-class variance

```
def otsu(img: np.array) -> np.array:
    """ Applies Otsu's binarization to the image. """
    h, n = histogram(img)
    X0, X1, p = 0, img.sum(), 0
    threshold, best = -1, X1*X1/n
    for t in range(256):
        u1, u0 = t*h[t], h[t]
        X0, X1, p = X0 + u1, X1 - u1, p + u0
        if p > 0 and n - p > 0:
            # for true intra-class variance
            # divide by -n then add E[X^2]
            var = X0*X0/p + X1*X1/(n - p)
            if var > best:
                threshold, best = t, var
    return __threshold(img, threshold)
```

Commentary

- Both approaches are $O(n)$ (assuming fixed 8-bit color)
- Both approaches generate identical thresholds (and images)
- However, intra-class variance is simpler and easier to derive
- It's also easier to convert to pure integer arithmetic, by storing fractions as numerator/denominator

Table of Contents

Example

Otsu's binarization ran on an image

(a) The original image.

(b) After Otsu's binarization.

Figure: Before and after Otsu's binarization.

Histogram and Inter-class Variance over Thresholds

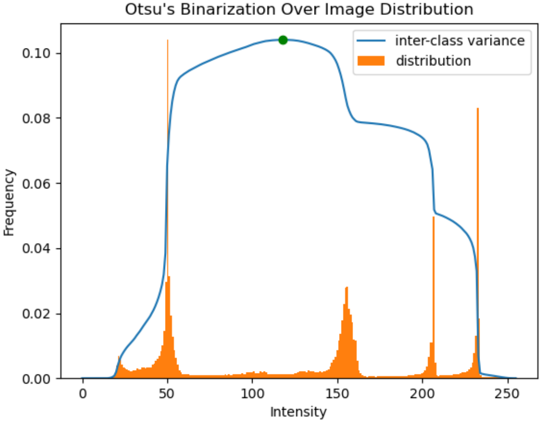

Figure: Inter-class variance over increasing threshold value.

Last Comments

- The image might have noise. Reduce noise with a Gaussian kernel, averaging, or other techniques.
- We've been using 8 bits per color channel for a pretty long time. But there's no reason why images with 16-bit color or even 48-bit color won't catch on. In that case Otsu's takes $O(N2^b)$ where b is the number of bits per channel, which will not scale with increasing color depth.
- People may need to quantize their images or apply a sophisticated 1D k-means algorithm, optimized for $k = 2$.

References

1. Dr. White's probability theory textbook
2. My lectures on
   - Probability theory
   - k-means
3. Wikipedia articles on
   - Probability axioms
   - Law of total probability
   - Law of total expectation
   - Law of total variance
   - k-means
   - Otsu's method
4. Otsu Thresholding
5. OpenCV documentation