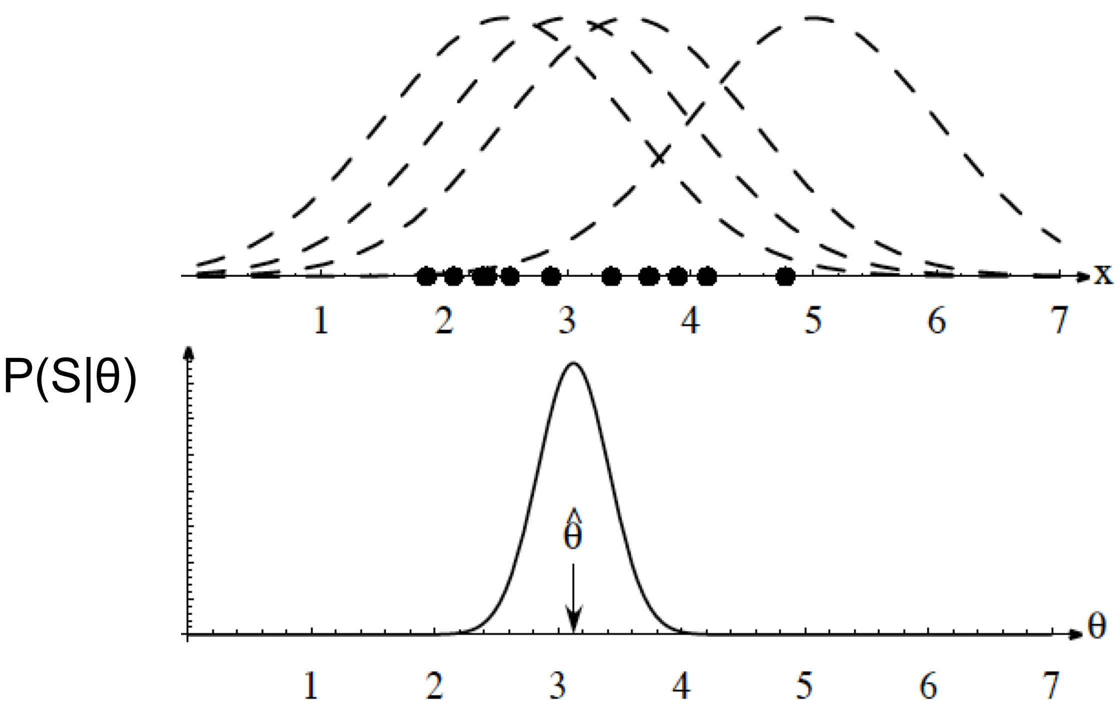
CMSC 510 – L08
Regularization Methods for Machine Learning

Instructor:
Dr. Tom Arodz

Recap: Maximum likelihood estimation

- Finding parameter θ that maximizes likelihood of seeing what we see in the training set S
  - Choose θ to maximize $P(S|\theta) = L(\theta \mid S) =$ *likelihood* of θ given dataset S
  - $L(\theta \mid S)=P(S|\theta) = \Pi_k P(x_k|\theta)$
- We assume θ for each class is fixed, but unknown to us
  - Some values of θ make the training set S more likely
  - Some values of θ make the training set S less likely



Simple example:
true mean of a normal distribution vs. average of some samples

We have samples S

How can we estimate the mean (θ)

Try all possible means, calculate probability of seeing the samples

Pick mean with highest probability

Recap: MLE/MAP vs Bayesian

- We're predicting something (some z) based on training set S and some new information u

  Law of total probability (we can condition on "weather in Iceland", or θ):

  - $p(z \mid S, u) = \Sigma_\theta \, p(z \mid \theta, S, u) \, p(\theta \mid S, u)$

  or in fact $= \int p(z \mid \theta, S, u) \, p(\theta \mid S, u) \, d\theta$

- Assume z and S are conditionally independent given θ, i.e., if we know θ, knowing also S doesn't change our knowledge of z
  - Then:
    - $p(z \mid S, u) = \Sigma_\theta \, p(z \mid \theta, u) \, p(\theta \mid S, u)$
- Assume also that $p(\theta \mid S, u) = p(\theta \mid S)$
  - the new info u alone (without z) does not impact our knowledge of θ
  - u may impact how we use θ, but that's in $p(z \mid \theta, u)$
- End result: $p(z \mid S, u) = \Sigma_\theta \, p(z \mid \theta, u) \, p(\theta \mid S)$
  - Find z that has highest probability given S and u

MLE/MAP is a simplification

$p(z \mid S, u) = \Sigma_\theta \, p(z \mid \theta, u) \, p(\theta \mid S)$

- Two options:
  - Maximum likelihood (MLE) / maximum a posteriori (MAP):
    - find single θ_{max} with highest $p(\theta \mid S)$,
      - θ_{max}=arg $\max_\theta P(\theta|S)$ i.e. θ_{max}=arg $\max_\theta P(S|\theta)P(\theta)$

$P(\theta|S)P(S) = P(S|\theta)P(\theta)$

    - Make an approximation:
      - approximate $p(\theta_{max} \mid S)$ as 1
      - all other θ had smaller $p(\theta \mid S)$, approximate $p(\theta_{other} \mid S)$ by 0
    - We end up with an approximation, but a much simpler formula for predictions/inference (for both MLE and MAP):
      - $p(z \mid S, u) = p(z \mid \theta_{max}, u)$
  - Bayesian estimation:
    - use formula $p(z \mid S, u) = \Sigma_\theta \, p(z \mid \theta, u) \, p(\theta \mid S)$
    - No approximations: we're using all possible values of θ,
      - weighing them by their probability given S

MLE/MAP vs Bayesian

- Bayesian estimation:
  - use formula $p(z \mid S, u) = \Sigma_\theta \, p(z \mid \theta, u)\, p(\theta \mid S)$
  - No approximations: we're using all possible values of θ,
    - weighing them by their probability given S
  - How to get $p(\theta \mid S)$
    - $p(\theta \mid S) = p(S \mid \theta)\, p(\theta) / P(S)$
    - $p(\theta \mid S) = p(S \mid \theta)\, p(\theta) / \Sigma_{\theta'}\, p(S \mid \theta')\, p(\theta')$
      - Or $p(\theta \mid S) = p(S \mid \theta) / \Sigma_{\theta'}\, p(S \mid \theta')$
        if no prior preference $p(\theta)$ – if all $p(\theta)$ are equal
    - We may get away without calculating $p_S = \Sigma_{\theta'}\, p(S \mid \theta')\, p(\theta')$
      - Depending on what we need $p(z \mid S, u)$ for
  - $p(z \mid S, u) = \Sigma_\theta \, p(z \mid \theta, u)\, p(S \mid \theta)\, p(\theta) / p_S$
    - vs. MLE/MAP: $p(z \mid S, u) = p(z \mid \theta_{max}, u)$

ML vs Bayesian estimation

- Known distribution shape, unknown parameters of the distribution, need to be estimated from data
- Maximum likelihood (MLE) approach:
  - We try to estimate single, most likely values of parameters from the training set
  - We use that single estimated value in all reasoning
- Bayesian learning approach:
  - We treat parameters as a random variable
  - We treat the training set as evidence that allows us to assign probabilities to different values of parameters
  - We use all possible parameter values, but each values carries different weight, its probability based on training set

ML vs Bayesian estimation

- Example ("*hidden dice*"):
  - we have positive integer samples
  - from a uniform distribution with an unknown maximum M
    - We only know M<=10
  - we observe set S of four values: 2, 4, 7, 8
  - what is p(x|S)?
- Maximum likelihood (ML) approach:
  - We choose M for which P(S|M)=P({2,4,7,8}|M) = P(2|M)*P(4|M)*P(7|M)*P(8|M) is highest
  - Based on that M, we assume p(x|S) = p(x|M)

  - What's the M?

ML vs Bayesian estimation

- Example:
  - we have positive integer samples
  - from a uniform distribution with an unknown maximum M
    - We only know $M<=10$
    - $P(x|M)=1/M$
  - we observe set S of four values: 2, 4, 7, 8
  - what is $p(x|S)$?
- Maximum likelihood (ML) approach:
  - We choose M that leads to highest value of:
    $P(S|M)=P(2|M)*P(4|M)*P(7|M)*P(8|M)$
  - If $M<8$ $P(S|M)=0$
  - If $M=8$ $P(S|M)=(1/8)^4$
  - If $M=9$ $P(S|M)=(1/9)^4$
  - If $M=10$ $P(S|M)=(1/10)^4$

ML vs Bayesian estimation

- Example:
  - we have positive integer samples
  - from a uniform distribution with an unknown maximum M
    - We only know $M<=10$
  - we observe set S of four values: 2, 4, 7, 8
  - what is $p(x|S)$?
- Maximum likelihood (ML) approach:
  - We choose $M=8$, and have: $p(x|S)=p(x|M=8)$
  - $P(x|S) = 1/8 = 0.125$ for any $x<=8$
    $= 0$ for $x=9$, $x=10$

ML vs Bayesian estimation

- Bayesian learning approach:
  - $p(x \mid S) = \int p(x \mid \theta_i)\, P(\theta_i \mid S)\, d\theta_i = \Sigma_i\, p(x \mid \theta_i)\, P(\theta_i \mid S)$

    $P(\theta_i \mid S) = P(S|\theta_i)P(\theta_i) \,/\, \Sigma_i\, P(S|\theta_i)P(\theta_i)$
    - Let's assume $P(M=1)=P(M=2)=\ldots=P(M=10)=0.1$
  - $P(S|\theta_i) = \Pi_k\, P(x_k|\theta_i)$
    - $P(x|S)= p(x|M=8)\, P(M=8|S) + p(x|M=9)\, P(M=9|S) + p(x|M=10)\, P(M=10|S)$
      - $P(M=8|S) = P(S|M=8)P(M=8) / \ldots = (1/8)^4 / [\, (1/8)^4+(1/9)^4+ (1/10)^4] =0.4916$
      - $P(M=9|S) = P(S|M=9)P(M=9) / \ldots = (1/9)^4 / [\, (1/8)^4+(1/9)^4+ (1/10)^4] =0.3069$
      - $P(M=10|S) = P(S|M=10)P(M=10) / \ldots = (1/10)^4 / [\, (1/8)^4+(1/9)^4+ (1/10)^4] =0.2014$
    - $P(x|S)= 0.49\, p(x|M=8) + 0.31\, p(x|M=9) + 0.20\, p(x|M=10)$
    - $P(x=9|S)= 0 + 0.31/9 + 0.20/10 = 0.054$ (not 0 as in ML)
    - $P(x=8|S)= 0.49/8 + 0.31/9 + 0.20/10 = 0.116$ (not 0.125)

ML vs Bayesian estimation

- Example:
  - we have positive integer samples
  - from a uniform distribution with an unknown maximum M
    - We only know M<=10
    - P(x|M)=1/M
  - we observe set S of four values: 2, 4, 7, 8
  - what is p(x|S)?
- Let's get more data:
  - We observed 4 more points (now we have 8)
  - Still, we haven't see 9 or 10,
  - the highest number we've seen is still 8
- How would that affect ML estimate of P(x|S)?
- How would that affect Bayesian estimate of P(x|S)?

ML vs Bayesian estimation

- Maximum likelihood (ML) approach:
  - We choose M that leads to highest value of:
    $$P(S|M)=P(2|M)*P(4|M)*P(7|M)*P(8|M)$$
    If $M<8$ $P(S|M)=0$
  - If $M=8$ $P(S|M)=(1/8)^8$
  - If $M=9$ $P(S|M)=(1/9)^8$
  - If $M=10$ $P(S|M)=(1/10)^8$

  - Still the same estimate:

  - $P(x|S)$ $= 1/8 = 0.125$ for any $x<=8$
    $= 0$ for $x=9$, $x=10$

ML vs Bayesian estimation

- Bayesian learning approach:
  - $p(x \mid S) = \int p(x \mid \theta_i)\, P(\theta_i \mid S)\, d\theta_i = \Sigma_i\, p(x \mid \theta_i)\, P(\theta_i \mid S)$

    $P(\theta_i \mid S) = P(S|\theta_i)P(\theta_i) \,/\, \Sigma_i\, P(S|\theta_i)P(\theta_i)$
    - Let's assume $P(M=1)=P(M=2)=\ldots=P(M=10)=0.1$
  - $P(S|\theta_i) = \Pi_k\, P(x_k|\theta_i)$
    - $P(x|S)= p(x|M=8)\, P(M=8|S) + p(x|M=9)\, P(M=9|S) + p(x|M=10)\, P(M=10|S)$
      - $P(M=8|S) = P(S|M=8)P(M=8) / \ldots$

        $= (1/8)^8 / [\, (1/8)^8+(1/9)^8+ (1/10)^8] = 0.6420$
      - $P(M=9|S) = P(S|M=9)P(M=9) / \ldots$

        $= (1/9)^8 / [\, (1/8)^8+(1/9)^8+ (1/10)^8] = 0.2502$
      - $P(M=10|S) = P(S|M=10)P(M=10) / \ldots$

        $= (1/10)^8 / [\, (1/8)^8+(1/9)^8+ (1/10)^8] = 0.1077$
    - $P(x|S)= 0.64\, p(x|M=8) + 0.25\, p(x|M=9) + 0.11\, p(x|M=10)$
    - $P(x=9|S)= 0 + 0.25/9 + 0.11/10 = 0.038$ (not 0 as in ML)
    - $P(x=8|S)= 0.65/8 + 0.25/9 + 0.11/10 = 0.118$ (not 0.125)

ML vs Bayesian estimation

- Bayesian learning approach:
  - $p(x \mid S) = \int p(x \mid \theta_i)\, P(\theta_i \mid S)\, d\theta_i = \Sigma_i\, p(x \mid \theta_i)\, P(\theta_i \mid S)$

    $P(\theta_i \mid S) = P(S|\theta_i)P(\theta_i) \;/\; \Sigma_i\; P(S|\theta_i)P(\theta_i)$
    - Let's assume $P(M=1)=P(M=2)=\ldots=P(M=10)=0.1$
  - $P(S|\theta_i) = \Pi_k\, P(x_k|\theta_i)$
    - $P(x|S)= p(x|M=8)\, P(M=8|S) + p(x|M=9)\, P(M=9|S) + p(x|M=10)\, P(M=10|S)$
      - $P(M=8|S) = P(S|M=8)P(M=8) / \ldots$
        $= (1/8)^4 / [\, (1/8)^4+(1/9)^4+ (1/10)^4] = 0.4916$
      - $P(M=9|S) = P(S|M=9)P(M=9) / \ldots$
        $= (1/9)^4 / [\, (1/8)^4+(1/9)^4+ (1/10)^4] = 0.3069$
      - $P(M=10|S) = P(S|M=10)P(M=10) / \ldots$
        $= (1/10)^4 / [\, (1/8)^4+(1/9)^4+ (1/10)^4] = 0.2014$
    - $P(x|S)= 0.49\, p(x|M=8) + 0.31\, p(x|M=9) + 0.20\, p(x|M=10)$
    - $P(x=9|S)= 0 + 0.31/9 + 0.20/10 = 0.054$ (not 0 as in ML)
    - $P(x=8|S)= 0.49/8 + 0.31/9 + 0.20/10 = 0.116$ (not 0.125)

Maximum likelihood estimation

- Maximum likelihood (ML):
  - Choose θ to maximize $L(\theta|S)=P(S|\theta) = \Pi_k P(x_k,\theta)$
  - Maximize *log-likelihood*: $\ln P(S|\theta) = \Sigma_k \ln P(x_k|\theta)$
- Maximum a posteriori (MAP):
  - Finding θ that maximizes $P(\theta|S) \sim P(S|\theta)P(\theta)$
  - maximize: $P(S|\theta)P(\theta) = \Pi_k P(x_k|\theta)P(\theta)$
  - Max.: $\ln P(S|\theta)P(\theta) = \Sigma_k \ln P(x_k|\theta) + \ln P(\theta)$
- In both versions:
  - We assume θ for each class is fixed, but unknown to us
  - We find the best single estimate of θ based on how a choice of θ influences S
  - We use θ for predictions

Large w and correlated features

- In real world:
  - measurement = signal + noise

- In real world:
  - very often we have features that are highly correlated

- For example:
  - neighboring pixels in an image
  - expression of genes that perform some function together

- **We want to avoid large weights!**
  - **But how?**

$P(S|\theta)$ – MLE – no preference for θ

$P(S|\theta)P(\theta)$ – MAP – has preference for θ

Maximum likelihood for P(y|x)

- Let's say we have a probability distribution P(y|x) of a certain shape, and the distribution is parameterized by a vector w, so P(y|x)=P(y|x,w)
- How to estimate w?
  - We have a training set S with m samples
  - We could do maximum likelihood estimation of w

$$\max P(y_1, y_2, \ldots, y_m \,|\, x_1, x_2, \ldots, x_m, w)$$

    - For which **w** are <u>all observed y's (set S_Y) for all corresponding x's (set S_X)</u> most likely?
    - We don't need to know anything about probability of x's
      - we're not estimating P(x,y), just P(y|x)
  - Samples are i.i.d: they're independent, so can we simplify this?

$$P(y_1, y_2, \ldots, y_m \,|\, x_1, x_2, \ldots, x_m, w)$$

Maximum likelihood for P(y|x)

- Let's say we have a probability distribution P(y|x) of a certain shape, and the distribution is parameterized by a vector w, so P(y|x)=P(y|x,w)
- How to estimate w?
  - We have a training set S with m samples
  - We could do maximum likelihood estimation of w

$$\max P(y_1, y_2, \ldots, y_m \,|\, x_1, x_2, \ldots, x_m, w)$$

    - For which **w** are the observed y's for x's most likely?
    - We don't need to know anything about probability of x's
      - we're not estimating P(x,y), just P(y|x)
  - Samples are i.i.d: they're independent, so:

$$P(y_1, y_2, \ldots, y_m \,|\, x_1, x_2, \ldots, x_m, w) = \Pi_{i=1}^{m} P_w(y_i \,|\, x_i, w)$$

  - ML estimate of **w:**

    w that maximizes: $\boxed{\max \Pi_{i=1}^{m} P(y_i \,|\, x_i, w)}$

Maximum likelihood

- We want: $p(y_i \mid x,S)$
  - P of y_i for sample x, given that we know training set $S=(S_Y,S_X)$
    - S_y: classes, S_X: features
  - We assume $P(y_i)$ comes from distrib. with parameter set w
    - Just one w needed, for P("+1") because: P("-1") = 1 - P("+1")
  - our knowledge learned from S (how class depends on x) will be encapsulated in a "good" w_s
    - $p(y \mid x,S) = p(y \mid x,w_s)$
  - How to get w_s ?
    - $w_s = \arg\max_w P(S_Y|S_X,w)$
    - $P(S_Y|S_X,w) = \Pi_k P(y_k|x_k,w)$
  - $w_s = \arg\max_w \Pi_k P(y_k|x_k,w)$

Logistic regression classifier

- Maximal likelihood estimation of parameter w of P(y|x,w) given the training set: *find w:*

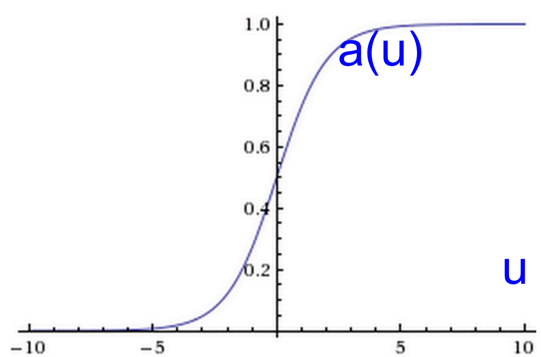
$$\max \Pi_{i=1}^{m} P(y_i \,|x_i, w\,)$$

- We need to assume some shape of P.
- Could it be that:

$$P(y_i \,|x_i, w\,) = a(y_i w^T x_i) = \frac{1}{1+\mathrm{e}^{-y_i w^T x_i}}$$

Is it mathematically ok?

What are the requirements for $a(yw^Tx)$ to be a valid probability distribution over possible classes, i.e., over set y={+1, -1} ?



Logistic regression classifier

- Maximal likelihood estimation of parameter w of P(y|x,w) given the training set: *find w:*

$$\max \Pi_{i=1}^{m} P(y_i \,|x_i, w\,)$$

- Could it be that:

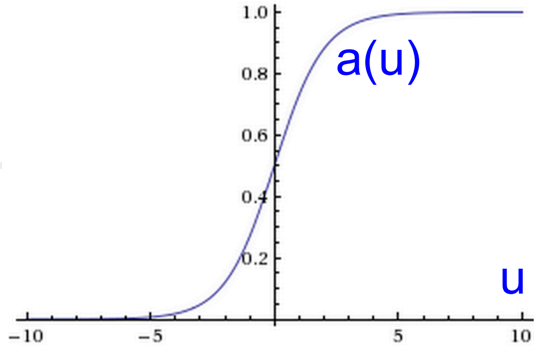
$$P(y_i \,|x_i, w\,) = a(y_i w^T x_i) = \frac{1}{1+\mathrm{e}^{-y_i w^T x_i}}$$

1. $a(yw^Tx)$ is between [0,1]
2. $a(+1*w^Tx)+a(-1*w^Tx)=1$

$$a(-u) = 1 - a(u)$$

  - So, a() could technically represent conditional probability of two classes, +1 and -1
  - But would it make any sense?
  - What's the shape of that conditional probability?

Logistic regression

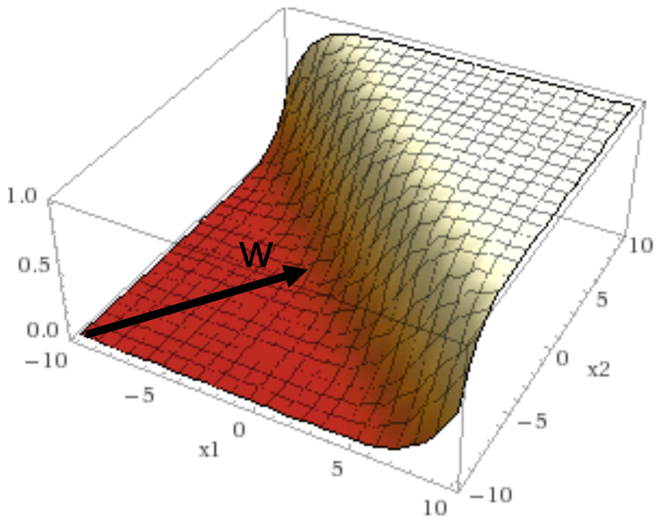


- Assumption: the distribution P(y|x,w) depends on vector of parameters **w** and is of the form:

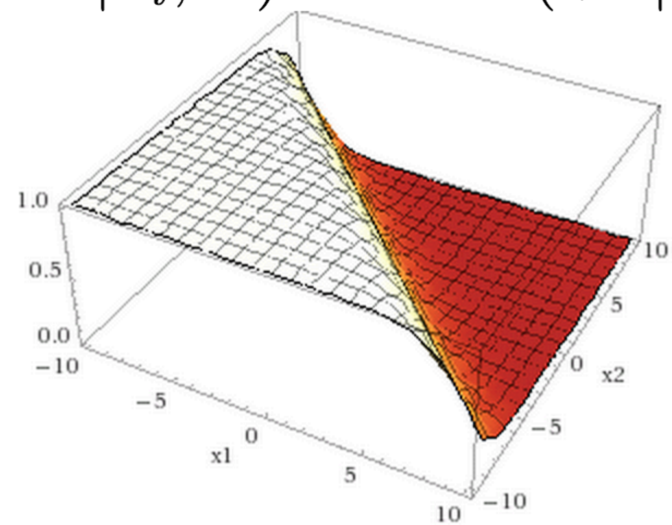
$$P(y_i\,|x_i, w) = a(y_i w^T x_i) = \frac{1}{1+e^{-y_i w^T x_i}}$$

- Example: w=[1,1], we have two features x_1, x_2

$P(+1\,|x_i, w)$



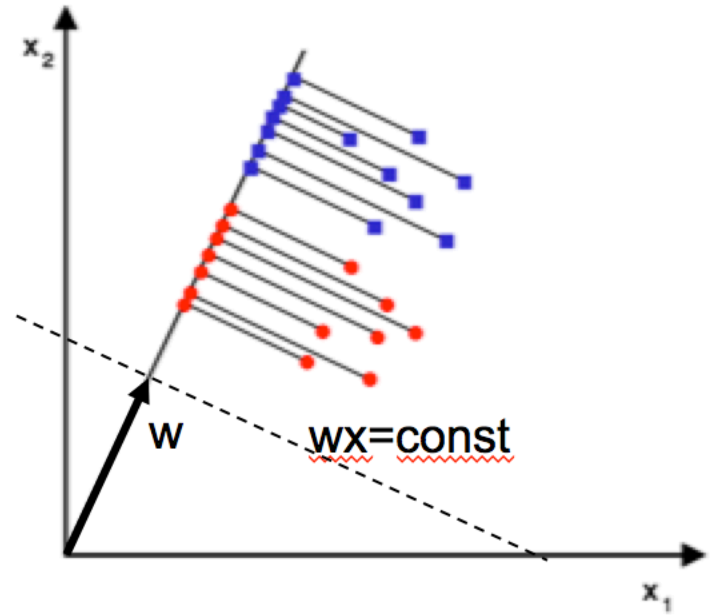
$P(-1\,|x_i, w) = 1 - P(+1\,|x_i, w)$

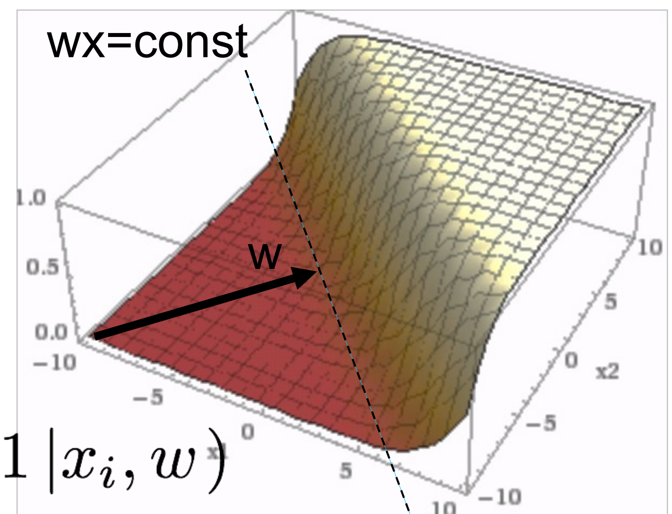


- This means: each class occupies a half-plane, and there's a straight region in the middle where classes overlap

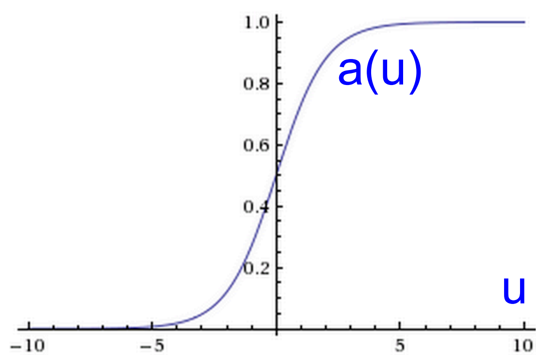
Logistic regression

- Similarity between $P(+1|x,w)$ and $a(u)$
  - $w^T x$ projects all samples on a single line (extension of vector w)
  - then a() is applied to values u on that line





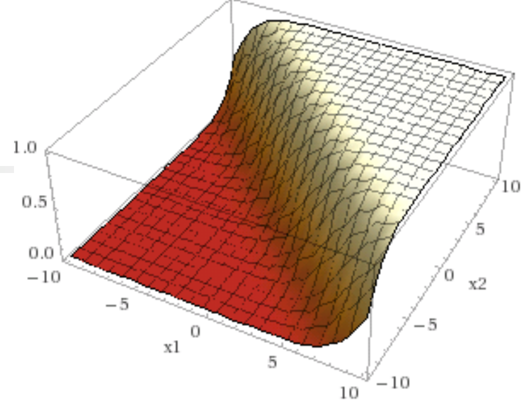
$P(+1\,|x_i, w)$

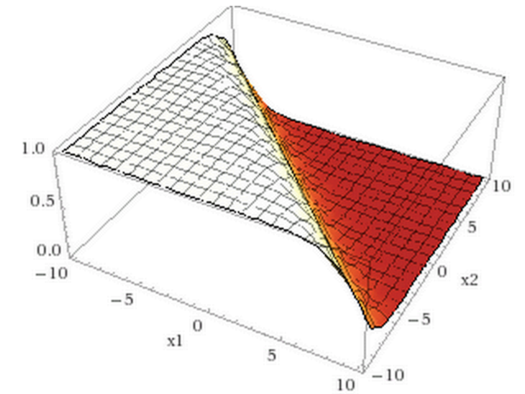


$$P(y_i\,|x_i, w) = a(y_i w^T x_i) = \frac{1}{1+e^{-y_i w^T x_i}}$$

Logistic regression

- Assumption: the distribution P(y|x,w) depends on vector of parameters **w** and is of the form:
  - each class occupies a half-plane,
    and there’s a band in the middle,
    around a straight line,
    where the two classes overlap



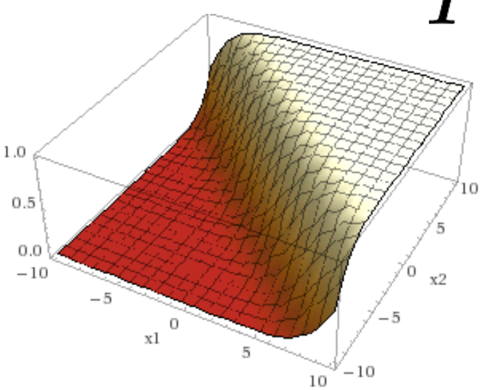


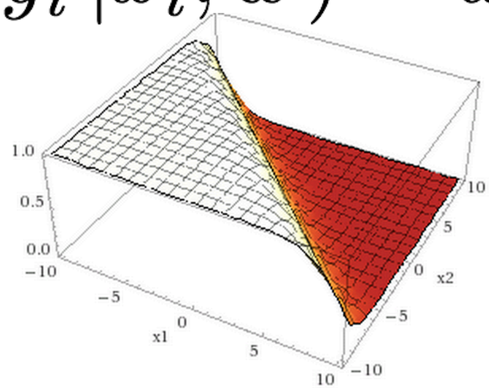
- It’s an assumption:
  for a given classification problem,
  it may be close to the truth
  or far from the truth

Logistic regression

- Under the assumption that class conditional probabilities depend on an unknown **w** in this way:

$$P(y_i\,|x_i,w\,) = a(y_i w^T x_i) = \frac{1}{1+\mathrm{e}^{-y_i w^T x_i}}$$





$$P(+1\,|x_i,w\,) \qquad P(-1\,|x_i,w\,) = 1 - P(+1\,|x_i,w\,)$$

- Maximum likelihood estimate of w is:

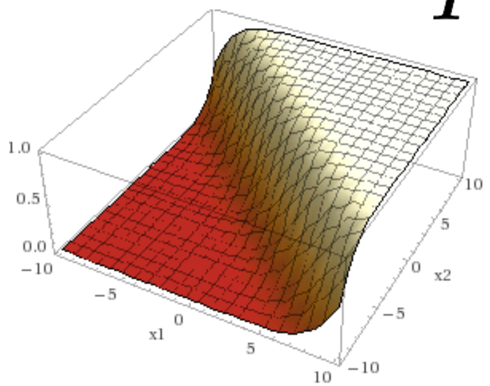
$$\max \Pi_{i=1}^{m} P(y_i\,|x_i,w\,)$$

- How to solve it?

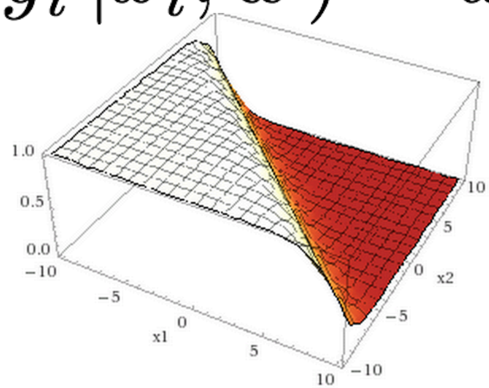
Logistic regression

- Under the assumption that class conditional probabilities depend on an unknown **w** in this way:

$$P(y_i\,|x_i, w) = a(y_i w^T x_i) = \frac{1}{1+\mathrm{e}^{-y_i w^T x_i}}$$



$P(+1\,|x_i, w)$



$P(-1\,|x_i, w) = 1 - P(+1\,|x_i, w)$

- Maximum likelihood estimate of w is:

$$\max \Pi_{i=1}^{m} P(y_i\,|x_i, w)$$

- Solve this instead:

$$\min -\frac{1}{m} \ln \Pi_{i=1}^{m} a(y_i w^T x_i)$$

- Or this:

$$\min \frac{1}{m} \sum_{i=1}^{m} \ln(1 + \mathrm{e}^{-y_i w^T x_i})$$